SuperCLUE
中文大模型综合性测评基准

中文大模型基准测评2025年5月报告

— 2025中文大模型阶段性进展5月评估

SuperCLUE团队

2025.05.28

SuperCLUE
中文大模型综合性测评基准
独立第三方AGI测评机构

精准量化通用人工智能（AGI）进展，定义人类迈向AGI的路线图

Accurately quantifying the progress of AGI,
defining the roadmap for humanity's journey towards AGI.

报告目录

报告摘要（一）

- **o4-mini(high)总分稳居第一，综合能力全面领先**

o4-mini(high)在本次5月测评中表现优异，总分达到70.51分，超过国内最好模型7.35分。该模型在推理、代码生成、智能体、指令遵循等多个方面表现出卓越的综合能力，特别是在代码生成（91.52）、指令遵循（68.07）方面得分较高。

- **国内推理模型崭露头角，部分领域优势突出**

Doubao-1.5-thinking-pro-205415、SenseNova V6 Reasoner等国内模型表现亮眼。其中，Doubao-1.5-thinking-pro-205415在文本创作与理解任务以81.04的高分领先其他模型。

- **国内大模型在指令遵循方面普遍低于海外模型**

Hunyuan-T1-20250403在国内模型中指令遵循得分第一，为36.97分，但是与海外模型指令遵循得分第一的o4-mini(high)相比，差距达到了31.1分，国内模型在指令遵循方面表现较弱，还有较大的提升空间。

- **小参数模型表现超出预期**

多款开源小参数量模型展现出惊人潜力。尤其是Qwen3系列，其中4B、8B和14B版本在推理任务上的分数均超过50分，超越了众多闭源大模型。

SuperCLUE模型象限（2025）

数据来源：SuperCLUE，2025年5月28日；

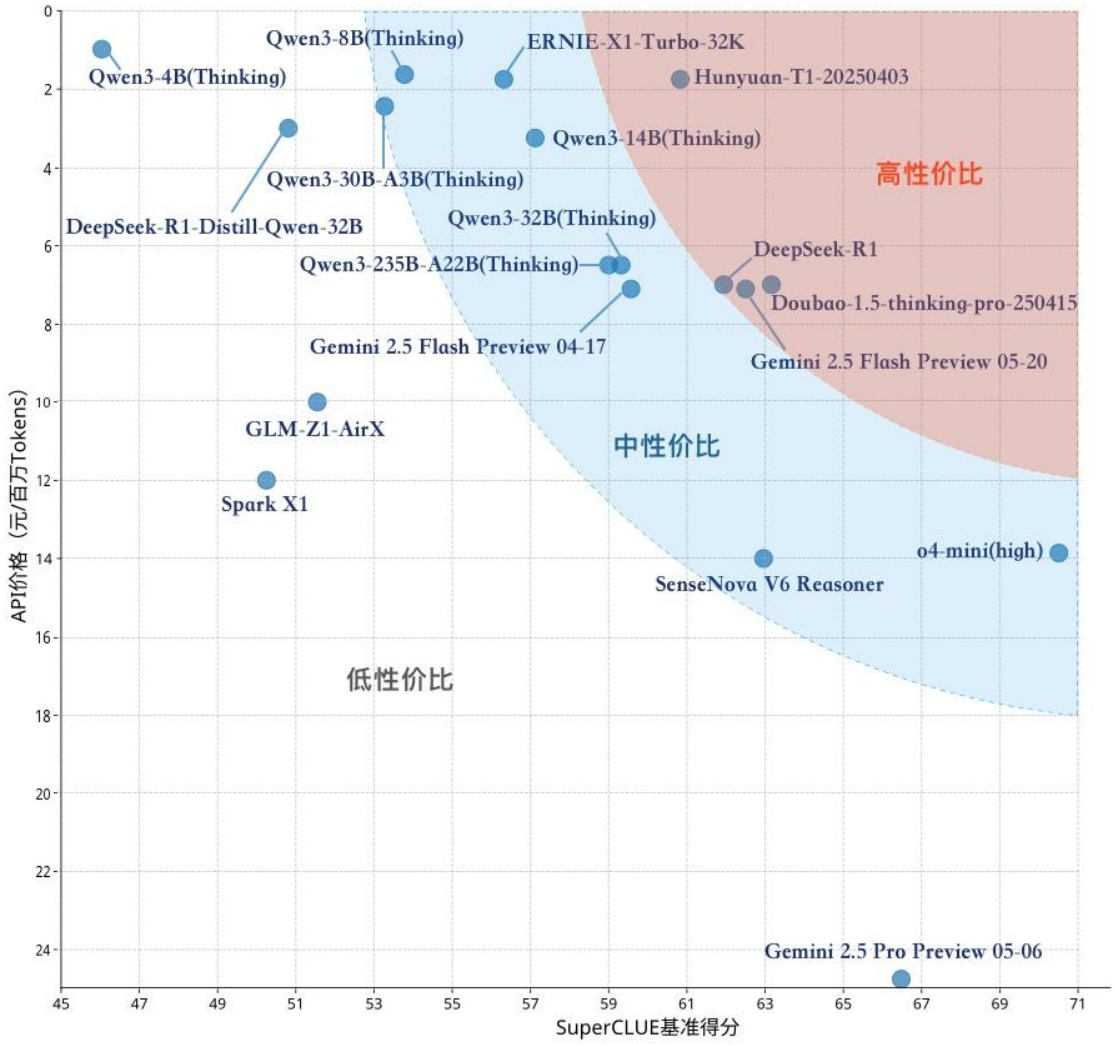
注：1. 两个维度的组成。推理能力包含：数学推理、科学推理、代码生成；应用能力包括：文本理解与创作、精确指令遵循、智能体Agent；

2. 四个象限的含义。它们代表大模型所处的不同阶段与定位，其中【潜力探索者】代表模型正在探索阶段未来拥有较大潜力；【技术领跑者】代表模型在基础技术方面具备领先性；【实用主义者】代表模型在场景应用深度上具备领先性；【卓越领导者】代表模型在基础和场景应用上处于领先位置，引领国内大模型发展。

报告摘要（二）

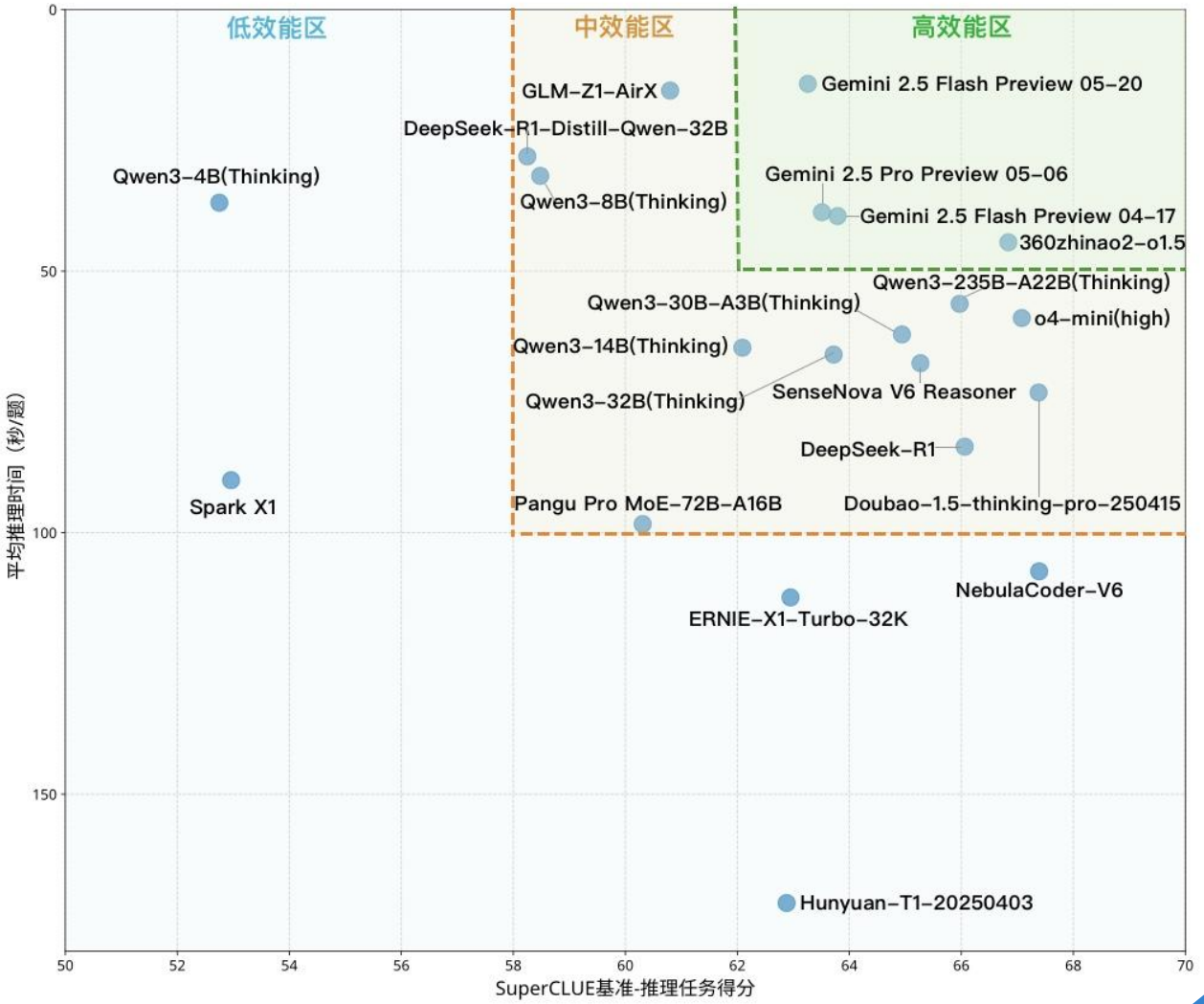
SuperCLUE
中文大模型综合性测评基准

大模型性价比区间分布



数据来源：SuperCLUE，2025年5月28日；推理任务得分为推理任务总分：数学推理、科学推理和代码的平均分。开源模型如Qwen3-32B(Thinking)使用方式为API，价格信息均来自官方信息。
注：部分模型API的价格是分别基于输入和输出的 tokens 数量确定的。这里我们依照输入 tokens 与输出 tokens 3:1 的比例来估算其整体价格。价格信息取自官方在5月的标准价格（非优惠价格）。

大模型推理效能区间分布



数据来源：SuperCLUE，2025年5月28日；
模型推理速度选取5月测评中具有公开API的模型。平均推理时间为所有测评数据推理时间的平均值（秒）。
推理任务得分为推理任务总分：数学推理、科学推理和代码生成的平均分。

第一部分

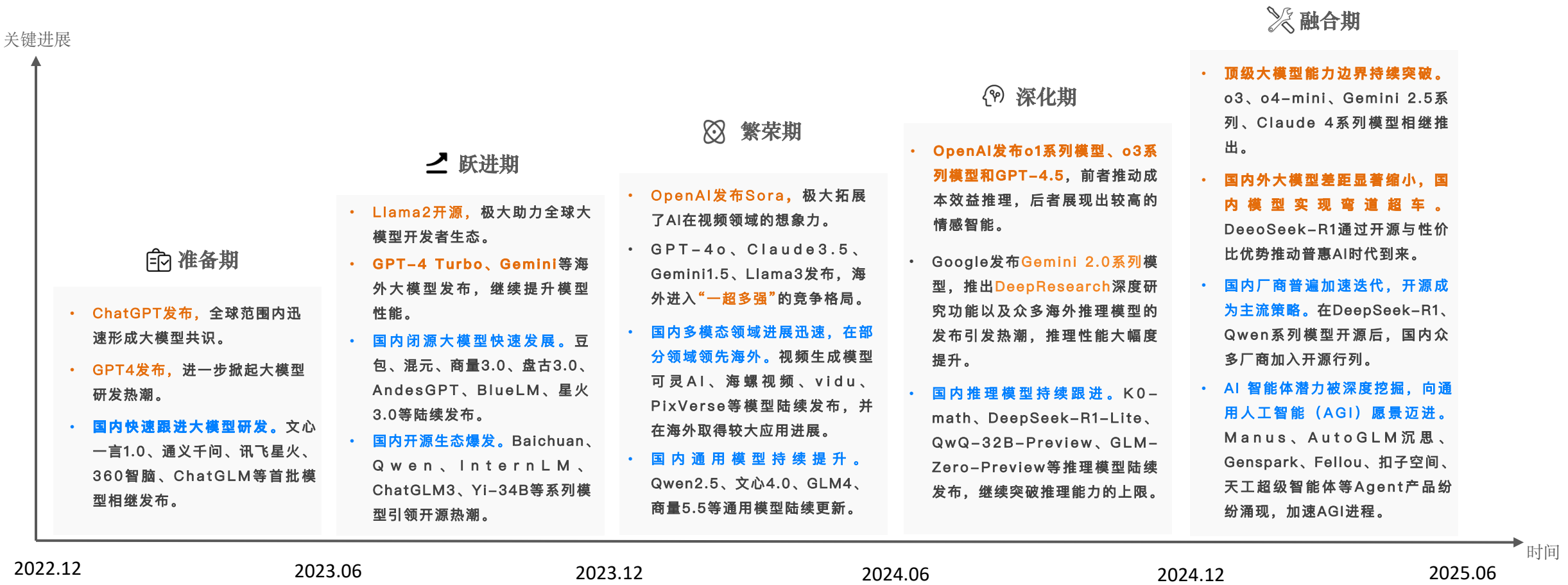
2025上半年度关键进展及趋势

1. 2025年上半年大模型关键进展
2. 2025年最值得关注的中文大模型全景图
3. 2025年国内外大模型差距

2025上半年大模型关键进展

◆自2022年11月30日ChatGPT发布以来，AI大模型在全球范围内掀起了有史以来规模最大的人工智能浪潮。国内外AI机构在过去2年半有了实质性的突破。具体可分为：准备期、跃进期、繁荣期、深化期和融合期。

SuperCLUE：AI大模型2025上半年关键进展



SuperCLUE：2025年最值得关注的中文大模型全景图

SuperCLUE
中文大模型综合性测评基准

文本

通用闭源：文心一言、通义千问、腾讯混元、商汤日日新 sensenova、BlueLM、360智脑、天工、MiLM、中科闻歌、紫东太初、澜舟科技 langboat、字节豆包、Kimi.ai、百川智能 BAICHUAN AI、MINIMAX、盘古大模型、云从科技 CLOUDWALK、DeepSeek、阶跃星辰、贝式计算 openbayes、Transn传神、智谱·AI、云知声 山海、零一万物、oppo、AndesGPT、ZTE中兴、讯飞星火、天翼AI、Scietrain 西湖心辰

通用开源：Qwen、DeepSeek、GLM-4、面壁小钢炮 MiniCPM、Yi、Hunyuan-Large、MiniMax-01、TeleChat2-35B、书生·浦语

推理：Qwen3、DeepSeek-R1、K1.5长思考、Step R-mini、360gpt2-o1.5、文心 X1 Turbo、GLM4 Z1系列、Hunyuan T1、Skywork o1

多模态

实时交互：星火极速、智谱清言、海螺AI、豆包、文小言、通义APP、日日新 5o sensenova、KIMI

图生视频：可灵AI、Vidu、PixVerse、WHEE、通义万相、阶跃视频、MINIMAX、清影

文生视频：可灵AI、即梦AI、清影、Vidu、PixVerse、海螺AI、HiDream.ai、通义万相

视觉理解：腾讯混元、阶跃星辰、Qwen2.5-VL、Doubao-vision、SenseChat-Vision、海螺AI、GLM-4v、书生·万象

文生图：即梦AI、混元-DiT、快手可图、CogView、讯飞星火、meitu、通义万相、文心一格

语音合成/声音复刻：Doubao-语音合成、百度TTS、讯飞语音合成、CosyVoice、Fish Audio、Speech-02

行业

医疗：百度灵医、医联MedGPT、百川AI全科医生、讯飞晓医

汽车：理想 MindGPT、极氪Kr大模型、DriveGPT、星睿AI、易车大模型

工业：奇智孔明AInno-15B、华为盘古工业大模型、羚羊工业大模型、SmartMore 思谋科技

金融：蚂蚁金融大模型、妙想金融大模型、度小满 轩辕大模型、HithinkGPT

教育：MathGPT、作业帮、子曰、豆包爱学、随时问、快对

营销：eclicktech 易点天下、探迹SalesGPT、句子互动 JUZI.BOT、BlueFocus

法律：Chat Law、CHINESE LAW 元典智库、得理 DLS 得理法搜、dandu.ai 案牍AI

其他：阅文集团 妙笔大模型、DP 深势分子大模型、……

智能体

通用闭源：AutoGLM 沉思、coze、MINIMAX、Fellou、心流、心响、纳米AI搜索 纳米AI超级搜索、flowith NEO、天工 天工超级智能体

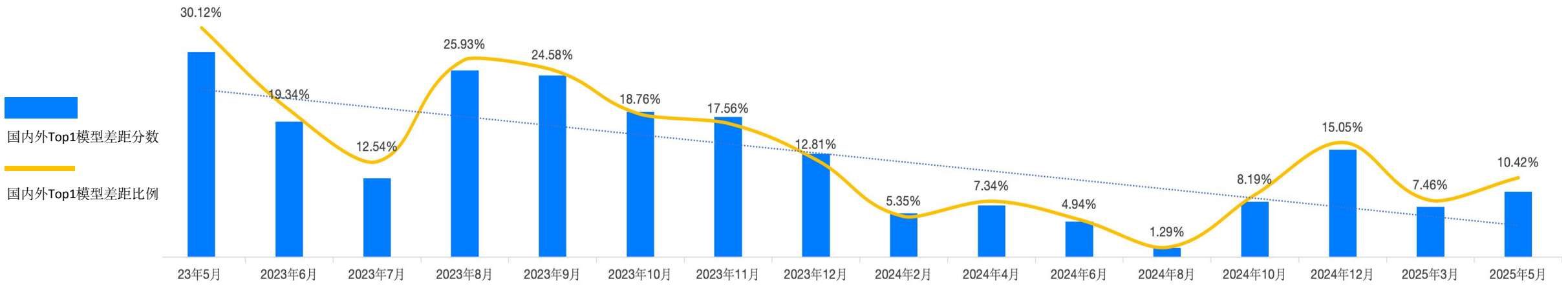
通用开源：OWL、OpenManus

深度研究 DeepResearch：智谱清言 沉思、纳米AI搜索、秘塔AI搜索、Qwen 深入研究、Reportify、问小白 小白研报

2025年国内外大模型差距

- **总体趋势上，国内外第一梯队大模型在中文领域的通用能力差距正在缩小。**2023年5月至今，国内外大模型能力持续发展。其中GPT系列模型为代表的海外最好模型经过了从GPT3.5、GPT4、GPT4-Turbo、GPT4o、o1、o3-mini、GPT-4.5、o3、o4-mini的多个版本的迭代升级。国内模型也经历了波澜壮阔的25个月的迭代周期。**但随着o4-mini的发布，差距从7.46%增加至10.42%。**

SuperCLUE基准：过去25个月国内外TOP大模型对比趋势



模型	23年5月	23年6月	23年7月	23年8月	23年9月	23年10月	23年11月	23年12月	24年2月	24年4月	24年6月	24年8月	24年10月	24年12月	25年3月	25年5月
GPT最新模型（GPT3.5、4、4-Turbo、4o、o1、o3-mini、GPT-4.5、o3、o4-mini）	76.67	78.76	70.89	81.03	83.20	87.08	89.79	90.63	92.71	79.13	81.00	79.67	75.85	80.4	76.01	70.51
国内TOP1	53.58	63.53	62.00	60.02	62.75	70.74	74.02	79.02	87.75	73.32	77.00	78.64	69.64	68.3	70.34	63.16
国内TOP2	49.52	62.58	59.35	55.70	62.61	70.42	72.88	76.54	86.77	72.58	76.00	76.24	69.00	68.3	66.38	62.96
国内TOP3	46.45	59.80	58.02	53.43	62.12	69.57	71.87	75.04	85.70	72.45	76.00	74.63	68.91	67.4	64.69	61.94

来源：SuperCLUE，2023年5月～2025年5月，期间发布的16次大模型基准测评报告。

第二部分

5月通用测评介绍

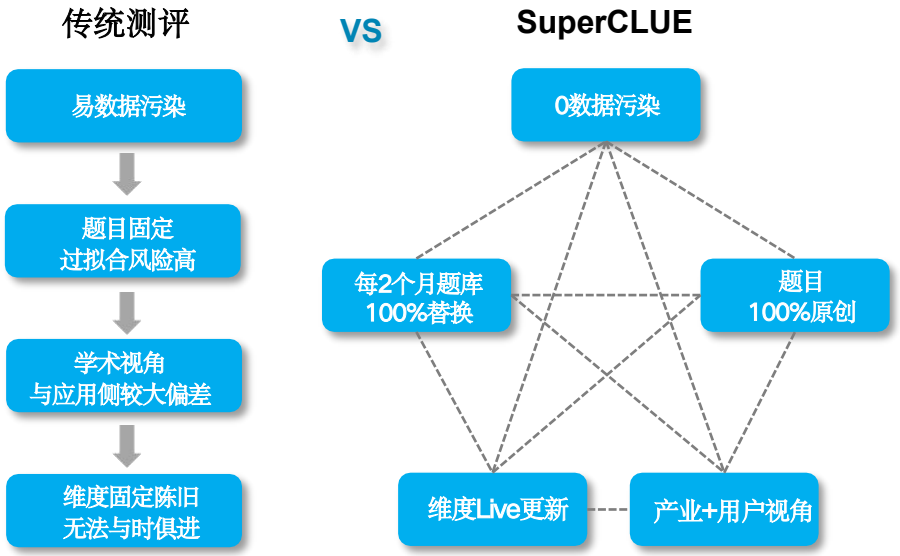
1. SuperCLUE基准介绍
2. SuperCLUE大模型综合测评体系
3. SuperCLUE通用测评基准数据集及评价方式
4. 各维度测评说明及示例
5. 测评模型列表

SuperCLUE基准介绍

SuperCLUE是大模型时代背景下CLUE基准的发展和延续，是独立、领先的通用大模型的综合性测评基准。中文语言理解测评基准CLUE（The Chinese Language Understanding Evaluation）**发起于2019年，**陆续推出过CLUE、FewCLUE、ZeroCLUE等广为引用的测评基准。

2019	2020	2021	2022	2023	2024.01-06	2024.07-12	2024.12-2025.03	2025.03-05
CLUE基准发布	发布多篇顶会论文，并承办了NLPCC 2020 开放测评任务	发布FewCLUE、DataCLUE、ZeroCLUE等多个知名测评基准	联合西湖大学发布DataCentricAI；发布SimCLUE大规模中文语义理解数据集	中文领域首个通用大模型测评基准SuperCLUE正式发布；发布首个中文多轮开放式测评基准SuperCLUE-Open；发布行业测评基准	发布数学、代码、长文本、RAG等测评基准；发布文生视频、文生图、视觉理解等多模态测评基准	发布AI编程助手等AI应用测评基准；发布链式推理等推理测评基准；发布多轮可执行智能体测评基准AgentCLUE	发布数学竞赛、科学推理、项目级代码生成等推理测评基准；发布实时音视频、视觉语言多模态测评基准；发布DeepSeek第三方平台性能基准	开源Math24o高中奥林匹克数学竞赛测评集；发布新版文生图、图生视频、视觉推理等中文测评基准；发布中文通用AI智能体测评基准AgentCLUE-General；发布中文大模型幻觉测评系列基准

SuperCLUE与传统测评的区别



SuperCLUE 三大特征

01 “Live”更新，0数据污染

测评题库每2个月100%替换且全部原创，杜绝过拟合风险。体系维度根据大模型进展Live更新。

02 测评方式与用户交互一致

测评方法与用户交互方式保持一致，测评任务贴近真实落地场景，高度还原用户视角。

03 独立第三方，无自家模型

完全独立的第三方评测机构，不研发自家模型。承诺提供无偏倚的客观、中立评测结果。

SuperCLUE大模型综合测评体系

基于大模型技术和应用发展趋势、以及基准测评专业经验，SuperCLUE构建出多领域、多层次的大模型综合性测评基准框架。从基础到应用覆盖：通用基准体系、文本专项系列基准、多模态系列基准、推理系列基准、Agent系列基准、AI应用基准、性能系列基准。为产业、学术和研究机构的大模型研发提供重要参考。

SuperCLUE大模型综合测评基准框架

通用基准

- 数学推理
- 科学推理
- 代码生成
- 智能体Agent
- 精确指令遵循
- 文本理解与创作

文本专项基准

幻觉

- 事实性幻觉 SuperCLUE-Fact
- 忠实性幻觉 SuperCLUE-Faith

- 长文本 SuperCLUE-Long
- 超长文本 SuperCLUE-200K
- 角色扮演 SuperCLUE-Role
- 检索增强 SuperCLUE-RAG
- 多轮对抗安全 SuperCLUE-Safety

多模态基准

视频

- 文生视频 SuperCLUE-T2V
- 图生视频 SuperCLUE-I2V

图像/视觉

- 实时音视频交互 SuperCLUE-Live
- 文生图 SuperCLUE-Image
- 多模态视觉语言 SuperCLUE-VLM

语音

- 实时语音交互 SuperCLUE-Voice
- 语音合成 SuperCLUE-TTS
- 声音复刻 SuperCLUE-Cloning

推理基准

- 视觉推理 SuperCLUE-VLR
- 科学推理 SuperCLUE-Science
- 全国高中数学竞赛 MathCLUE
- 项目级代码 SuperCLUE-Project
- 链式推理 SuperCLUE-COT
- 小学奥数 SuperCLUE-Math6o
- 代码 SuperCLUE-Code3
- 数学多步推理 SuperCLUE-Math6
- 竞赛级别代码
- 研究生级别数学

Agent基准

通用

- 可执行智能体 AgentCLUE
- 不可执行智能体 SuperCLUE-Agent
- 中文通用AI智能体 AgentCLUE-General

终端智能体

- 手机
- PC
- 眼镜
- ...

行业智能体

- 零售
- 教育
- 医疗
- ...

应用基准

AI产品

- AI搜索 SuperCLUE-AISearch
- 代码助手 SuperCLUE-Coder
- 实时音视频 SuperCLUE-Live

行业

- 金融 SuperCLUE-Fin
- 工业 SuperCLUE-Industry
- 汽车 SuperCLUE-Auto
- 智能座舱 SuperCLUE-Icabin

性能基准

DeepSeek-R1系列

- DeepSeek-R1 第三方联网搜索能力测试（网页端）
- DeepSeek-R1 第三方稳定性测试（App端）
- DeepSeek-R1 第三方稳定性测试（API端）
- DeepSeek-R1 第三方稳定性测试（网页端）
- 大模型推理速度测评

其余系列

- 开源数据集 Math24o
- 深度研究 DeepResearch

已发布　即将发布

注：通用基准介绍可在报告中查看，其余基准可点击对应链接跳转至最新的发布文章。

SuperCLUE通用测评基准数据集及评价方式

本次2025年5月报告聚焦通用能力测评，由六大维度构成。题目均为**原创新题**，总量为1579道多轮简答题。

【SuperCLUE通用数据集】共有数学推理、科学推理、代码生成、智能体Agent、精确指令遵循、文本理解与创作六大任务；

【SuperCLUE评价方式】分为基于人工校验参考答案的评估（0-1得分）、基于代码单元测试的评估（0-1得分）、结合任务完成与否、系统状态比对的评估（0-1得分）、基于规则脚本的评估（0-1得分）、人工校验参考答案的、多维度评价标准的评估。

SuperCLUE通用基准数据集及评价方式

1.数学推理

介绍：主要考察模型运用数学概念和逻辑进行多步推理和问题解答的能力。包括但不限于几何学、代数学、概率论与数理统计等竞赛级别数据集。

评价方式：*基于人工校验参考答案的评估（0-1得分）*

2.科学推理

介绍：主要考察模型在跨学科背景下理解和推导因果关系的能力。包括物理、化学、生物等在内的研究生级别科学数据集。

评价方式：*基于人工校验参考答案的评估（0-1得分）*

3.代码生成

介绍：主要考察模型在处理编程任务时理解和生成代码的能力。HumanEval的中文升级版，涵盖数据结构、基础算法、数学问题、数据科学等多种类型的代码数据集。

评价方式：*基于代码单元测试的评估（0-1得分）*

4.智能体Agent

介绍：主要考察在中文场景下基于可执行的环境，LLM作为执行代理，在多轮对话中调用工具完成任务的能力。包括两大任务类型：常规单轮对话和常规多轮对话。

评价方式：*结合任务完成与否、系统状态比对的评估（0-1得分）*

5.精确指令遵循

介绍：主要考察模型的指令遵循能力，包括但不限于定义的输出格式或标准来生成响应，精确地呈现要求的数据和信息。

评价方式：*基于规则脚本的评估（0-1得分）*

6.文本理解与创作

介绍：主要考察模型在处理文本相关任务时的理解和创作能力。包括但不限于文本摘要、阅读理解、指代消解、长文本等基础语义理解和生成创作数据集。

评价方式：*人工校验参考答案的、多维度评价标准的评估*

SuperCLUE-数学推理评测数据集

数学推理

主要考察模型运用数学概念和逻辑进行多步推理和问题解答的能力。包括但不限于几何学、代数学、概率论与数理统计等竞赛级别数据集。

测评方法

评测流程：

我们采用模型评价的方式来评估大模型在【数学推理】任务上的表现，使用高级模型（如gpt4o）判断其回答中的最终答案和参考答案是否相符，以确认大模型在一道问题上的正确率（正确或错误，即0或1）。

以【代数学】任务为例，假设在一个具体的题目中，模型A的最终答案与参考答案等价，则该题目的具体得分为1分；反之，假设在另一个具体的题目中，模型A的最终答案与参考答案不等价，则该题目的具体得分为0分。

最后，将所有题目的得分进行归一化处理，以确保满分为100分。这样最终的得分可以清晰地反映模型在【数学推理】任务中的表现。这一方法实现了完全的客观评估性。

示例-数学推理

问题： 在三维欧几里得空间中，四面体$ABCD$的顶点A、B、C、D满足以下条件：

1. 向量$\overrightarrow{AB}$与$\overrightarrow{AC}$的夹角为60°，且$\overrightarrow{AB}\cdot\overrightarrow{AC}=3$；
2. 向量$\overrightarrow{AB}$与$\overrightarrow{AD}$的夹角为60°，且$\overrightarrow{AB}\cdot\overrightarrow{AD}=5$；
3. 向量$\overrightarrow{AC}$与$\overrightarrow{AD}$的夹角为60°，且$\overrightarrow{AC}\cdot\overrightarrow{AD}=7$。

求四面体$ABCD$的体积。 请将最终的答案放在\boxed{}中。

模型回答： 要计算四面体$ABCD$的体积，我们可以利用向量的混合积。

设从顶点A出发的向量$\overrightarrow{AB}=\mathbf{a}$，$\overrightarrow{AC}=\mathbf{b}$，$\overrightarrow{AD}=\mathbf{c}$，

则体积V为这三个向量混合积绝对值的六分之一，

即 $V=\frac{1}{6}|\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})|$

步骤1：确定向量的模长 已知向量间的夹角均为60°，

<篇幅有限省略>

因此，混合积的绝对值为$\sqrt{420}=2\sqrt{105}$。

步骤3：计算体积 体积为： $V=\frac{1}{6}\times 2\sqrt{105}=\frac{\sqrt{105}}{3}$

最终答案：$\boxed{\frac{\sqrt{105}}{3}}$

评价打分： 问题及改进的反馈意见：

模型的回答清晰地展示了计算四面体体积的步骤，包括使用向量的混合积公式、点积关系、格拉姆行列式计算等。回答逻辑清晰，步骤详尽，最终结论与参考答案一致。没有明显的计算错误或逻辑问题。

打分与评价开始：

1. 基于标准1的<结论准确性>：

模型的最终答案为$\boxed{\frac{\sqrt{105}}{3}}$，与参考答案$\frac{\sqrt{105}}{3}$完全一致。回答准确，符合标准1的要求。得分：1分。

综合以上，AI助手的回答的综合得分（平均分）为[[1.00]]。

SuperCLUE-科学推理评测数据集

科学推理

主要考察模型在跨学科背景下理解和推导因果关系的能力。包括物理、化学、生物等在内的研究生级别科学数据集。

测评方法

评测流程：

我们采用模型评价的方式来评估大模型在【科学推理】任务上的表现，使用高级模型（如gpt4o）判断其回答中的最终答案和参考答案是否相符，以确认大模型在一道问题上的正确率（正确或错误，即0或1）。

以【物理】任务为例，假设在一个具体的题目中，模型A的最终答案与参考答案等价，则该题目的具体得分为1分；反之，假设在另一个具体的题目中，模型A的最终答案与参考答案不等价，则该题目的具体得分为0分。

最后，将所有题目的得分进行归一化处理，以确保满分为100分。这样最终的得分可以清晰地反映模型在【科学推理】任务中的表现。这一方法实现了完全的客观评估性。

示例-科学推理

问题： 在 298 K 时，某可逆电池的标准电动势 E° 为 1.10 V，其温度系数 $\left(\frac{\partial E}{\partial T}\right)_p$ 为 $-4.0 \times 10^{-4}\ \mathrm{V/K}$。若该电池在 323 K 时，两电极反应物浓度分别为 $[\mathrm{Cu}^{2+}] = 0.01\ \mathrm{M}$，$[\mathrm{Zn}^{2+}] = 0.1\ \mathrm{M}$（假设活度系数 $\gamma \approx 1$），气体常数 $R = 8.314\ \mathrm{J/(mol\cdot K)}$，法拉第常数 $F = 96485\ \mathrm{C/mol}$。试计算该电池在 323 K 时的电动势 E（单位：V），结果保留三位小数。

模型回答：

1. **计算 323 K 时的标准电动势 $E^\circ(323\ \mathrm{K})$**

   利用温度系数与电动势的关系：$E^\circ(T_2) = E^\circ(T_1) + \left(\frac{\partial E}{\partial T}\right)_p \cdot (T_2 - T_1)$

   代入数据：$E^\circ(323\ \mathrm{K}) = 1.10\ \mathrm{V} + (-4.0 \times 10^{-4}\ \mathrm{V/K}) \cdot (323 - 298)\ \mathrm{K} = 1.10 - 0.01 = 1.09\ \mathrm{V}$

   <篇幅有限省略>

   计算对数项：$\ln 10 \approx 2.3026$，$\frac{8.314 \times 323}{2 \times 96485} \approx \frac{2685.422}{192970} \approx 0.01391$

   因此：$E = 1.09 - 0.01391 \times 2.3026 \approx 1.09 - 0.0321 = 1.0579\ \mathrm{V}$

   保留三位小数得：$E \approx \boxed{1.058\ \mathrm{V}}$

评价打分：

问题及改进的反馈意见：

AI助手的回答逻辑清晰，步骤分明，计算过程详细且准确。它正确地应用了公式和数据，最终得出的结果与参考答案一致。没有明显的不足。

打分与评价开始：

1. **基于标准1的<结论准确性>；**

AI助手的回答最终结果为1.058，与参考答案完全一致，且计算过程无误。基于标准1的得分为1分。

综合以上，AI助手的回答的综合得分（平均分）为[[1.00]]。

SuperCLUE-代码生成评测数据集

代码生成

主要考察模型在处理编程任务时理解和生成代码的能力。HumanEval的中文升级版，涵盖数据结构、基础算法、数学问题、数据科学等多种类型的代码数据集。

测评方法

评测流程：

整体测评流程包括：**1.获取模型答案；2.提取功能函数；3.测试功能函数；4.计算模型得分。**

1.获取模型答案：

使用特定的prompt要求模型按照特定格式回答，以方便后续提取。对于一个代码补全问题，每个模型获取一次答案。

2.提取功能函数：

获取到模型回复之后，结合prompt要求和模型回答设定规则提取函数代码。

3.测试功能函数：

对于一个代码补全问题，将提取得到的代码和该问题对应的N个测试用例组装成N个可独立运行的单元测试程序，在沙箱环境运行（沙箱环境python版本设置为3.10）。

4.计算模型得分：

对于一个代码补全问题和1个测试用例，构成一个单元测试，通过测试得1分。最终计算模型在所有单元测试中得分平均分*100作为最终的得分。

示例-代码生成

问题：

```
import itertools
from typing import List, Any, Dict

def candidate(programmers: List[Dict[str, Any]], required_skills: List[str], team_size_k: int) -> List[int]:
    """
    组建一个由恰好 `team_size_k` 名成员组成的编程竞赛团队。
<篇幅有限省略>
    输出:
    - 选定团队成员的 'id' 列表，按升序排列。如果无解，则为空列表。
    """
```

模型回答：

```
import itertools
from typing import List, Any, Dict

def candidate(programmers: List[Dict[str, Any]], required_skills: List[str], team_size_k: int) -> List[int]:
    n = len(programmers)
    best_synergy = None
    best_ids: List[int] = []
    required = set(required_skills)
<篇幅有限省略>
        elif total == best_synergy:
            ids = sorted(programmers[i]['id'] for i in comb)
            if ids < best_ids:
                best_ids = ids
    return best_ids if best_synergy is not None else []
```

测试功能函数：

```
def check(candidate):
    assert candidate([{'id': 0, 'skills': ['Python', 'AI'], 'synergy_scores': [0,5,2]}, {'id': 1, 'skills': ['Java', 'Web'], 'synergy_scores': [5,0,3]}, {'id': 2, 'skills': ['Python', 'DB'], 'synergy_scores': [2,3,0]}], ['Python', 'Web'], 2) == [0, 1]
    assert candidate([{'id':0,'skills':['P'],'synergy_scores':[0,10]},{'id':1,'skills':['J'],'synergy_scores':[10, 0]}],['P','J'],2) == [0,1]
    assert candidate([{'id':0,'skills':['P'],'synergy_scores':[0]}],['P'],2) == []
    assert candidate([], ['P'], 1) == []
```

计算模型得分： 测试通过，得1分。

各维度测评说明及示例——智能体Agent测评说明

SuperCLUE-智能体Agent数据集

智能体Agent

主要考察在中文场景下基于可执行的环境，LLM作为执行代理，在多轮对话中调用工具完成任务的能力。包括两大任务类型：常规单轮对话和常规多轮对话。

测评方法

评测流程：

每一题包含N轮对话，每一轮对话为一个记分单元，答对记1分。

我们将模型回复的函数调用进行实际执行后判定是否回复正确。评价维度包括：函数调用执行结果、执行后的系统状态。

对于每一轮模型回复的评价方法：

1. 获取模型答案之后，从中提取可执行的函数调用。
2. 执行提取的函数调用得到执行结果集1，以及执行后的系统状态1；
3. 执行人类标注的ground truth(函数调用)得到执行结果集2，系统状态2；
4. 将执行结果集1与执行结果集2进行比较，系统状态1和系统状态2进行比较；

如果执行结果集2是执行结果集1的子集，系统状态1和系统状态2相同则认为模型的本轮回复正确，得1分。

最后，计算答对的记分单元的比例，转化为百分制得到模型的总分。即：

模型得分=（答对的记分单元数量/记分单元总数量）*100

示例-智能体Agent

描述：本次Agent测评共包含10个API，本题用到一个API。

车辆控制系统：该工具属于车辆控制系统，允许用户控制汽车的各个方面，如发动机、车门、空调、灯光等。

prompt：你是一个组合函数完成任务的专家。你将会收到一个问题和一些可以调用的函数及其描述。基于问题，你需要调用一个或者多个函数来实现问题要求。
对于问题中的计算需求，你必须调用提供的函数进行计算，否则视为错误。
如果使用你收到的函数无法完成问题要求，请指出。如果你收到的问题缺乏调用函数所必要的参数<篇幅有限省略>
下面提供了一个列表，列表中是一些你可以调用的函数以及函数的功能和参数描述，使用json格式<篇幅有限省略>

【问题】：我想要知道北京和西安之间的距离。

【模型回复1】：【函数调用开始】```[get_zipcode_based_on_city(city="北京"), get_zipcode_based_on_city(city="西安")]```【函数调用结束】

【模型得到的反馈1】：下面是根据你回复的函数调用执行得到的结果,你可以据此判断是否完成了用户的需求。如果已经完成，你可以停止回复函数调用，根据执行结果回答用户的问题。如果没有完成任务，你可以决定给出新的函数调用。
下面是函数调用结果：
第1个函数调用：get_zipcode_based_on_city(city='北京'), 执行返回结果为：{"zipcode": "352415"}
第2个函数调用：get_zipcode_based_on_city(city='西安'), 执行返回结果为：{"zipcode": "541227"}
注意：一旦你决定进行函数调用，你的回复中只能包括函数调用。你的函数调用（function calls）必须严格遵循以下格式给出："【函数调用开始】```[func_name1(params_name1=params_value1, params_name2=params_value2...), func_name2(params)]```【函数调用结束】"，不应该包括任何其他文本。

【模型回复2】：【函数调用开始】```[estimate_distance(cityA="352415", cityB="541227")]```【函数调用结束】

【模型得到的反馈2】：下面是根据你回复的函数调用执行得到的结果,你可以据此判断是否完成了用户的需求。如果已经完成，你可以停止回复函数调用，根据执行结果回答用户的问题。如果没有完成任务，你可以决定给出新的函数调用。
下面是函数调用结果：
第1个函数调用：estimate_distance(cityA='352415', cityB='541227'), 执行返回结果为：{"distance": 780.0}
注意：一旦你决定进行函数调用，你的回复中只能包括函数调用。你的函数调用（function calls）必须严格遵循以下格式给出："【函数调用开始】```[func_name1(params_name1=params_value1, params_name2=params_value2...), func_name2(params)]```【函数调用结束】"，不应该包括任何其他文本。

【模型回复3】：北京和西安之间的距离约为780公里。

本题共1轮问题，解决问题需要调用3次工具，模型的回答中3次工具调用均正确，则本轮回答正确，记1分。

各维度测评说明及示例——精确指令遵循测评说明

SuperCLUE-精确指令遵循数据集

精确指令遵循

主要考察模型的指令遵循能力，包括但不限于定义的输出格式或标准来生成响应，精确地呈现要求的数据和信息。

测评方法

评测流程：

我们完成了一个针对不同的语言模型的（文本输出）精确指令遵循表现的测试，根据设定的评估标准进行评估。评估的结果为布尔值（True 或 False）。

以【写一篇关于可再生能源对减少碳排放的作用的短文。要求文章不得使用“二氧化碳”这个词，字数不少于10个字，以JSON格式返回】任务为例：

设定的评价标准包括：遵循关键词限制、符合字数要求、输出格式正确。模型只有在命中所有指令的条件下会得到True的返回结果。

例如：

模型输出结果1：{ "response": "可再生能源在减少碳排放和减轻气候变化方面起着关键作用，未来应加快推广。" }

符合所有要求，返回True。

模型输出结果2：{ "response": "使用可再生能源可以减少二氧化碳排放。" }

违背了关键词限制，返回False。

最后，我们将所有题目的评估结果进行汇总，根据模型完全命中的题目数量计算模型的最终命中率，命中率以百分比为单位。具体来说，命中率计算公式为：

命中率 = (模型完全命中的题目数量 ÷ 总题目数量) * 100%

示例-精确指令遵循

问题： 近年来，随着全球化的加速和国际贸易的扩展，港口物流在现代经济中的地位日益重要。港口不仅是货物进出口的主要节点，也是地区经济发展的动力。现代化港口包含多种设施，如货运码头、仓储区域、集装箱处理设备等，同时依赖复杂的信息系统进行管理。然而，港口物流也面临诸多挑战，包括环保压力、交通拥堵和技术更新需求。为了解这些挑战并寻找解决方案，我们需要深入研究港口物流的各个方面，包括技术应用、管理模式和未来发展趋势。
请撰写一篇关于“港口物流未来技术趋势”的分析文章。**在文章中必须出现“自动化”、“绿色能源”和“信息化”这三个关键词，每个关键词必须至少出现3次。文章只能有4个段落（不要出现标题），每个段落用“~~”符号分隔。第3段必须恰好由5个句子（句子用中文句号或中文问号或中文感叹号分隔）组成。整篇文章只能有12个句子。**请确保内容结构清晰，并对未来的技术趋势进行合理推测。

模型回答： 近年来，港口物流在全球供应链中扮演着越来越重要的角色，推动了区域经济的快速发展。面对日益严峻的环保要求，越来越多港口开始引入绿色能源项目以减少碳排放。在此过程中自动化技术和信息化平台成为提升作业效率的关键。
~~
自动化堆场、智能导航车辆和无人驾驶设备正在改造传统码头作业流程，缩短了周转时间并降低人力成本。同时，借助信息化手段实时监控货物流向，提高了资源利用率。
~~
智能港口将通过自动化设备实现全天候无人值守操作，从而大幅提升生产效率和作业安全。大规模部署的绿色能源系统如太阳能光伏和风力发电正在为港口设施提供可靠的清洁动力，降低对化石燃料的依赖。信息化平台与人工智能算法深度融合能够优化航线调度、货柜追踪及能耗管理，释放更多数据价值。5G网络和物联网技术进一步支撑了港口内部的实时数据交换和远程控制，有助于精准预测需求和资源配置。未来几年，随着区块链技术的引入和大数据分析的普及，港口物流的透明度和协作水平将进入新的高度。
~~
面向未来，港口将持续融合自动化、信息化与绿色能源解决方案，朝着智能化、低碳和高效的方向迈进。各方需加强技术研发与跨领域协同，才能确保港口物流系统的可持续发展和竞争优势。

评价打分：
本题存在三个精确指令考察：1）关键词出现频率；2）段落数量；3）句子数量。
模型的回答符合要求，得一分。

SuperCLUE-文本理解与创作评测数据集

文本理解与创作

主要考察模型在处理文本相关任务时的理解和创作能力。包括但不限于文本摘要、阅读理解、指代消解、长文本等基础语义理解和生成创作数据集。

测评方法

评测流程：

我们采用高级AI模型（例如GPT-4o）来评估不同任务的表现，根据设定的评价标准进行打分（1-5分）。

以【生成与创作】任务为例，设定的评价标准包括：符合基本要求、语言与逻辑连贯性、推理链条完整度。

假设在一个具体的题目中，模型A的得分如下：符合基本要求3分、语言与逻辑连贯性4分、推理链条完整度5分。首先，我们计算这一题目的平均得分：

$$\frac{3+4+5}{3} = 4$$

因此，该题目的具体得分为4分。

最后，将所有题目的得分进行归一化处理，以确保满分为100分。这样，每个任务的得分都能清晰地反映模型在不同标准下的表现。

示例-文本理解与创作

问题： 请总结以下文本，形成摘要：
我正经过清风桥，脚下的石板路面带着一种古旧的温润感，被无数脚步打磨得光滑，却又在缝隙里藏着青苔的湿意与岁月的低语。桥身并非全然的平坦，带着一丝不易察觉的弧度向上微微拱起，仿佛是大地温和的呼吸，引领着行人走向河流的中央，也走向视野的开阔处。栏杆是铁艺的，缠绕着复古的卷草纹饰，油漆早已斑驳，露出底下暗红色的铁锈，像是一道道时光刻下的伤痕，却也平添了几分沧桑的韵味。我将手轻轻搭在冰凉的栏杆上，指尖能感受到那细微的震动，或许是桥下水流冲击桥墩的回响，又或许是远处车流驶过桥面时传递而来的生命脉动。风，确实如桥名一般，带着清爽的气息，从宽阔的江面上吹拂而来，撩动我的发梢，也试图吹散心中那一点莫名的滞重感。<篇幅有限省略>

模型回答： 作者漫步清风桥，细致描绘了桥的古旧韵味（光滑石板、微拱桥身、斑驳铁艺栏杆）、周边自然景致（碧绿江水、碎金般的阳光、薄雾中的城市天际线、掠过的水鸟）及两岸对比（现代滨江公园的生活气息与古旧民居的烟火味）。桥上行人百态（下棋老人、嬉闹孩童、写生学生等）与作者的驻足观察交织，触发对时光流转、生活复杂的感悟——桥不仅是地理连接，更承载着时间的过往与未来、情感的此岸与彼岸。作者由此追忆少年时的憧憬，感慨当下心境的平和，在江风与流水的陪伴中，获得心灵的澄明与前行的坚定，最终带着桥上的宁静思考，重新融入生活。

评价打分：
打分与评价开始：
1. 基于标准1的符合基本要求：
AI助手准确回应了用户的问题，并围绕文本内容进行了高度相关的总结。总结结构清晰，满足用户的提问意图与需求。略微可以强化对某些细节的表达与隐含情感的深入梳理。得分：4.5分。

2. 基于标准2的事实正确性：
AI助手整体回答的事实性较强，与原文本内容一致，涵盖了关键部分如桥上的细节描写、两岸的对比、行人百态与心灵感悟等。未发现明显的事实错误，且表达符合文本意境要求。得分：5.0分。

3. 基于标准3的清晰度：
回答语言表达流畅且清晰，条理明晰，阅读起来易于理解。回答虽简洁，但对于部分意象的表现力略显不足，无法完全展现原文本感性的细腻美。可以补充一些具体的描述增强表现力。得分：4.0分。

综合以上，AI助手的回答的综合得分（平均分）为[[4.50]]。

测评模型列表

本次测评数据选取了**SuperCLUE-5月**测评结果，模型选取了国内外有代表性的**43个大模型在5月份的版本**。

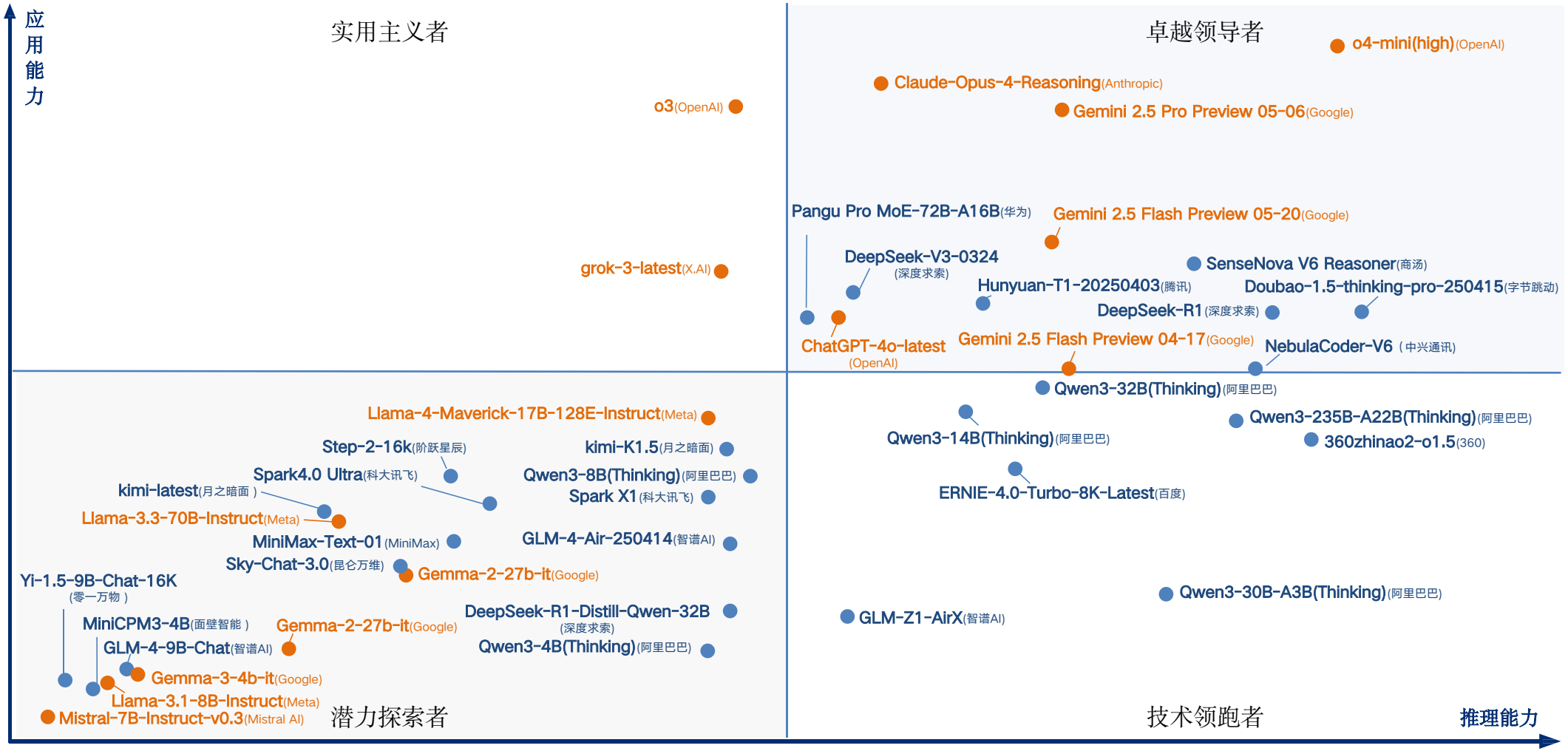
模型	机构	简介	模型	机构	简介
1.Qwen3-235B-A22B(Thinking)	阿里巴巴	官方发布的MoE推理模型，使用阿里云公开的API：qwen3-235b-a22b。	24.Gemma-3-27b-it	Google	Gemma3开源系列27B模型，使用官方API：gemma-3-27b-it。
2.Qwen3-30B-A3B(Thinking)	阿里巴巴	官方发布的MoE推理模型，使用阿里云公开的API：qwen3-30b-a3b。	25.Gemma-3-12b-it	Google	Gemma3开源系列12B模型，使用官方API：gemma-3-12b-it。
3.Qwen3-32B(Thinking)	阿里巴巴	官方发布的Dense推理模型，使用阿里云公开的API：qwen3-32b。	26.Gemma-3-4b-it	Google	Gemma3开源系列4B模型，使用官方API：gemma-3-4b-it。
4.Qwen3-14B(Thinking)	阿里巴巴	官方发布的Dense推理模型，使用阿里云公开的API：qwen3-14b。	27.Llama-4-Maverick-17B-128E-Instruct-FP8	Meta	使用together.ai的接口：meta-llama/Llama-4-Maverick-17B-128E-Instruct-FP8。
5.Qwen3-8B(Thinking)	阿里巴巴	官方发布的Dense推理模型，使用阿里云公开的API：qwen3-8b。	28.Llama-3.3-70B-Instruct	Meta	使用together.ai的接口：meta-llama/Llama-3.3-70B-Instruct-Turbo。
6.Qwen3-4B(Thinking)	阿里巴巴	官方发布的Dense推理模型，使用阿里云公开的API：qwen3-4b。	29.Llama-3.1-8B-Instruct	Meta	使用together.ai的接口：meta-llama/Meta-Llama-3.1-8B-Instruct-Turbo。
7.Step-2-16k	阶跃星辰	官方公开发布的API版本：step-2-16k。	30.ChatGPT-4o-latest	OpenAI	与ChatGPT上的GPT-4o同版本，对应OpenAI官方的API名称:chatgpt-4o-latest。
8.Sky-Chat-3.0	昆仑万维	昆仑万维发布的千亿级别 MOE 模型，使用官方的API接口。	31.o4-mini(high)	OpenAI	使用方式为AZURE OpenAI Service的API接口，reasoning_effort参数设置为：high。
9.GLM-4-Air-250414	智谱AI	官方发布的全新语言模型，使用官方的API：GLM-4-Air-250414。	32.o3	OpenAI	OpenAI在2025年4月16日发布的最新推理模型o3，使用方式为POE：o3。
10.GLM-Z1-AirX	智谱AI	官方发布的GLM-Z1系列推理模型，使用官方的API：GLM-Z1-AirX。	33.360zhinao2-o1.5	360	官方提供的小范围内测版本，使用方式为API。
11.GLM4_9B_Chat	智谱AI	官方开源的GLM-4-9B-Chat，对应huggingface 仓库名称：THUDM/glm-4-9b-chat。	34.grok-3-latest	X.AI	X.AI在2025年2月19日推出的模型版本，使用官方API，版本名称为：grok-3-latest。
12.DeepSeek-V3-0324	深度求索	深度求索在2025年3月24日发布的新版本V3模型，使用官方API：deepseek-chat。	35.ERNIE-X1-Turbo-32K	百度	百度发布的深度思考模型，使用百度千帆的API，版本名称为：ernie-x1-turbo-32k。
13.DeepSeek-R1	深度求索	深度求索在2025年1月20日发布的开源推理模型，使用官方API：deepseek-reasoner。	36.Pangu Pro MoE-72B-A16B	华为	官方提供的小范围内测版本，使用方式为API。见技术报告：https://arxiv.org/pdf/2505.21411。
14.DeepSeek-R1-Distill-Qwen-32B	深度求索	基于Qwen2.5-32B的蒸馏模型，使用阿里云API：deepseek-r1-distill-qwen-32b。	37.NebulaCoder-V6	中兴通讯	官方提供的小范围内测版本，使用方式为API。
15.Spark X1	科大讯飞	科大讯飞发布的API版本：Spark X1。	38.MiniMax-Text-01	MiniMax	官方发布的新一代模型，使用方式为API，版本名称为：MiniMax-Text-01。
16.Spark4.0 Ultra	科大讯飞	科大讯飞发布的API版本：Spark4.0 Ultra。	39.Hunyuan-T1-20250403	腾讯	官方发布的深度思考模型，使用方式为API：hunyuan-t1-20250403。
17.kimi-latest	月之暗面	与Kimi智能助手产品使用的大模型同版本，使用API：kimi-latest。	40.SenseNova V6 Reasoner	商汤	官方提供的小范围内测版本，使用方式为API。
18.kimi-K1.5	月之暗面	月之暗面推出的推理模型，使用官网网页（开启“K1.5长思考”模式）。	41.Yi-1.5-9B-Chat-16K	零一万物	官方开源版本，huggingface 仓库名称：01-ai/Yi-1.5-9B-Chat-16K。
19.Doubao-1.5-thinking-pro-250415	字节跳动	官方发布的深度思考模型，使用方式为API：doubao-1-5-thinking-pro-250415。	42.Mistral-7B-Instruct-v0.3	Mistral AI	官方开源版本，对应huggingface 仓库名称：mistralai/Mistral-7B-Instruct-v0.3。
20.Claude-Opus-4-Reasoning	Anthropic	官方发布的Claude Opus 4，使用方式为POE：Claude-Opus-4-Reasoning。	43.MiniCPM3-4B	面壁智能	官方开源版本。对应huggingface 仓库名称：openbmb/MiniCPM3-4B。
21.Gemini 2.5 Pro Preview 05-06	Google	Gemini 2.5 Pro的预览版本，使用官方API：gemini-2.5-pro-preview-05-06。	/	/	/
22.Gemini 2.5 Flash Preview 04-17	Google	Gemini 2.5 Flash的预览版本，使用官方API：gemini-2.5-flash-preview-04-17。	/	/	/
23.Gemini 2.5 Flash Preview 05-20	Google	Gemini 2.5 Flash的预览版本，使用官方API：gemini-2.5-flash-preview-05-20。	/	/	/

第三部分

总体测评结果与分析

1. SuperCLUE模型象限
2. SuperCLUE通用能力测评榜单
3. SuperCLUE-Agent：智能体测评分析
4. SuperCLUE性价比区间分布
5. SuperCLUE大模型综合效能区间分布
6. 国内大模型成熟度-SC成熟度指数
7. 评测与人类一致性验证
8. 开源模型榜单
9. 10B级别小模型榜单
10. 端侧5B级别小模型榜单

SuperCLUE模型象限（2025）

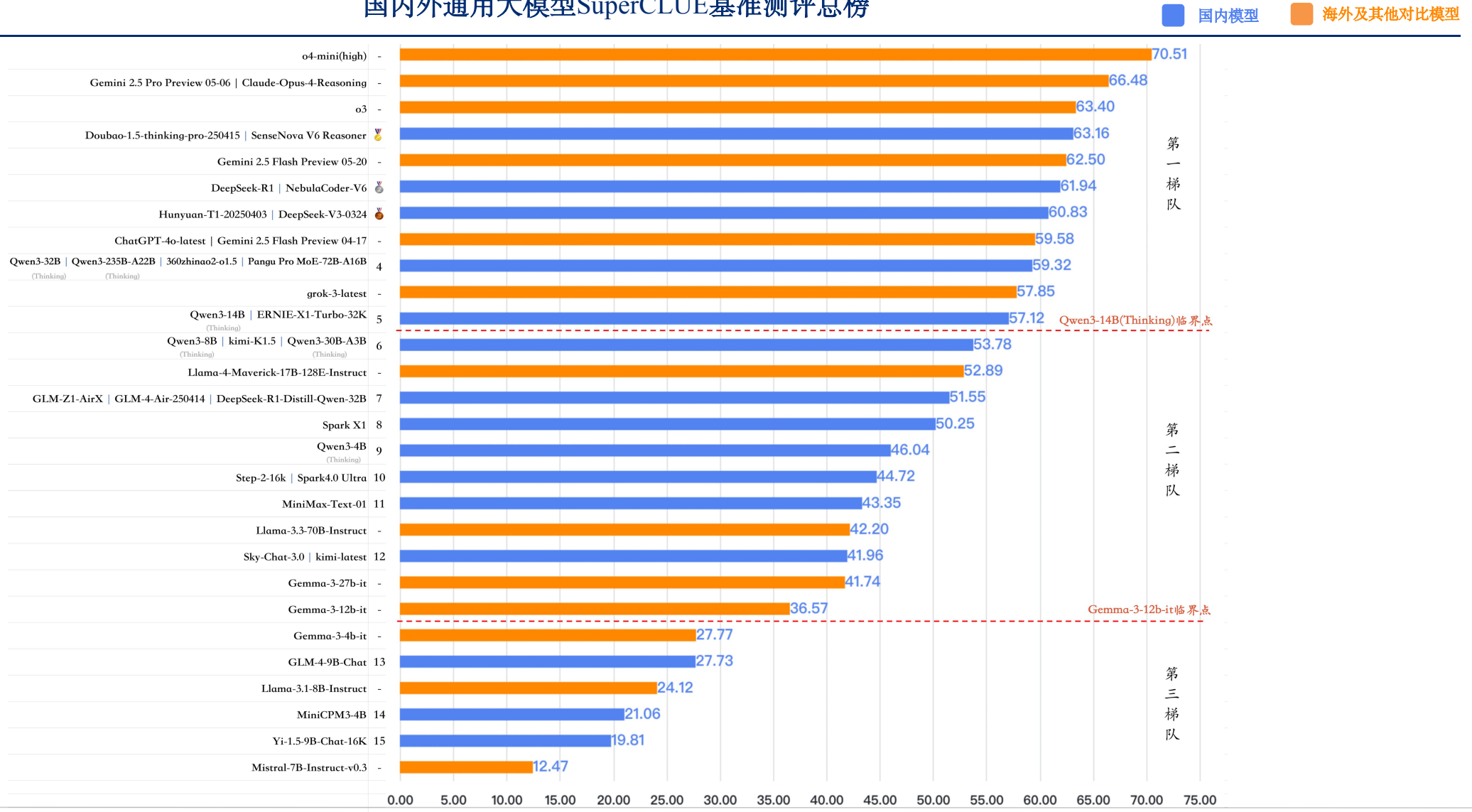


来源：SuperCLUE，2025年5月28日；

注：1. 两个维度的组成。推理能力包含：数学推理、科学推理、代码；应用能力包括：文本理解与创作、指令遵循、Agent能力；

2. 四个象限的含义。它们代表大模型所处的不同阶段与定位，其中【潜力探索者】代表模型正在探索阶段未来拥有较大潜力；【技术领跑者】代表模型在基础技术方面具备领先性；【实用主义者】代表模型在场景应用深度上具备领先性；【卓越领导者】代表模型在基础和场景应用上处于领先位置，引领国内大模型发展。

国内外通用大模型SuperCLUE基准测评总榜



来源：SuperCLUE，2025年5月28日；

注：由于部分模型分数较为接近，为了减少问题波动对排名的影响，本次测评将相距1分区间的模型定义为并列，报告中分数展示为并列中高分。海外模型仅对比参考不参与排名。

SuperCLUE-总榜

SuperCLUE测评基准2025年5月总体表现

排名	模型名称	机构	开源/闭源	总分	数学推理	科学推理	代码生成	智能体Agent	精确指令遵循	文本理解与创作	使用方式
-	o4-mini(high)	OpenAI	闭源	***70.51***	57.72	52.00	91.52	76.01	68.07	77.75	API
-	Gemini 2.5 Pro Preview 05-06	Google	闭源	***66.48***	55.65	43.56	91.32	74.39	53.78	80.15	API
-	Claude-Opus-4-Reasoning	Anthropic	闭源	***65.74***	51.61	45.54	86.98	70.61	59.10	80.62	POE
-	o3	OpenAI	闭源	***63.40***	56.45	25.74	88.76	62.84	66.95	79.67	POE
1	Doubao-1.5-thinking-pro-250415	字节跳动	闭源	***63.16***	62.10	***52.08***	***87.97***	60.47	***35.29***	***81.04***	API
1	SenseNova V6 Reasoner	商汤	闭源	***62.96***	62.10	47.52	86.19	***69.59***	***31.93***	80.43	API
-	Gemini 2.5 Flash Preview 05-20	Google	闭源	***62.50***	56.45	46.53	86.79	64.85	41.18	79.18	API
2	DeepSeek-R1	深度求索	开源	***61.94***	62.10	49.50	86.59	***62.50***	31.37	79.57	API
2	NebulaCoder-V6	中兴通讯	闭源	***61.59***	***62.39***	***53.19***	86.59	59.46	27.45	80.47	API
3	Hunyuan-T1-20250403	腾讯	闭源	***60.83***	57.50	47.52	83.63	58.78	***36.97***	***80.59***	API
3	DeepSeek-V3-0324	深度求索	开源	***60.10***	55.65	41.58	84.81	***68.24***	30.25	80.04	API
-	ChatGPT-4o-latest	OpenAI	闭源	***59.58***	49.19	42.08	90.34	67.57	29.13	79.16	API
-	Gemini 2.5 Flash Preview 04-17	Google	闭源	***59.57***	57.26	47.52	86.59	52.03	34.73	79.26	API
4	Qwen3-32B(Thinking)	阿里巴巴	开源	***59.32***	56.45	47.52	***87.18***	58.45	26.61	79.70	API
4	Qwen3-235B-A22B(Thinking)	阿里巴巴	开源	***59.00***	58.87	48.51	***90.53***	51.01	26.05	79.03	API
4	360zhinao2-o1.5	360	闭源	***58.90***	***62.86***	50.52	87.14	50.00	23.53	79.36	API
4	Pangu Pro MoE-72B-A16B	华为	开源	***58.75***	59.68	39.60	81.66	***62.50***	28.57	***80.48***	API
-	grok-3-latest	X.AI	闭源	***57.85***	49.19	34.65	83.04	65.88	34.45	79.90	API
5	Qwen3-14B(Thinking)	阿里巴巴	开源	***57.12***	57.26	***51.49***	77.51	52.36	24.09	80.03	API
5	ERNIE-X1-Turbo-32K	百度	闭源	***56.32***	57.63	48.98	82.25	50.34	19.33	79.37	API
6	Qwen3-8B(Thinking)	阿里巴巴	开源	***53.78***	60.48	42.57	72.39	45.27	22.97	79.00	API
6	kimi-K1.5	月之暗面	闭源	***53.72***	59.68	38.61	73.37	50.68	25.77	74.18	网页
6	Qwen3-30B-A3B(Thinking)	阿里巴巴	开源	***53.27***	61.29	50.50	83.04	18.58	26.33	79.85	API
-	Llama-4-Maverick-17B-128E-Instruct	Meta	开源	***52.89***	44.35	38.61	75.15	54.05	30.81	74.39	API
7	GLM-Z1-AirX	智谱AI	开源	***51.55***	***64.41***	46.00	71.99	44.59	13.17	69.11	API
7	GLM-4-Air-250414	智谱AI	开源	***51.45***	53.23	35.64	77.91	43.58	21.85	76.49	API
7	DeepSeek-R1-Distill-Qwen-32B	深度求索	开源	***50.81***	60.33	38.89	75.54	35.47	19.38	75.22	API
8	Spark X1	科大讯飞	闭源	***50.25***	53.23	31.68	73.96	39.86	23.25	79.54	API
9	Qwen3-4B(Thinking)	阿里巴巴	开源	***46.04***	50.81	33.66	73.77	17.91	21.57	78.50	API
10	Step-2-16k	阶跃星辰	闭源	***44.72***	30.65	18.81	70.81	49.66	19.61	78.76	API
10	Spark4.0 Ultra	科大讯飞	闭源	***44.63***	36.29	22.77	67.46	44.01	19.89	77.38	API
11	MiniMax-Text-01	MiniMax	开源	***43.35***	29.03	20.79	71.20	43.92	19.33	75.80	API
-	Llama-3.3-70B-Instruct	Meta	开源	***42.20***	30.65	12.87	70.22	41.22	27.73	70.51	API
12	Sky-Chat-3.0	昆仑万维	闭源	***41.96***	37.10	19.80	60.16	35.81	22.97	75.91	API
12	kimi-latest	月之暗面	闭源	***41.91***	24.19	13.86	71.99	51.69	15.69	74.04	API
-	Gemma-3-27b-it	Google	开源	***41.74***	32.26	17.82	68.05	34.12	20.45	77.72	API
-	Gemma-3-12b-it	Google	开源	***36.57***	27.42	4.95	68.44	26.69	15.97	75.96	API
-	Gemma-3-4b-it	Google	开源	***27.77***	16.94	2.97	48.92	11.82	11.52	74.47	API
13	GLM-4-9B-Chat	智谱AI	开源	***27.73***	16.94	6.93	43.98	18.24	7.84	72.47	模型
-	Llama-3.1-8B-Instruct	Meta	开源	***24.12***	9.68	4.95	39.25	16.22	8.96	65.64	API
14	MiniCPM3-4B	面壁智能	开源	***21.06***	8.87	4.95	34.91	1.35	7.02	69.28	模型
15	Yi-1.5-9B-Chat-16K	零一万物	开源	***19.81***	6.45	2.97	20.71	12.16	7.56	69.01	模型
-	Mistral-7B-Instruct-v0.3	Mistral AI	开源	***12.47***	5.65	0.99	9.07	1.01	3.36	54.76	模型

注：数据来源SuperCLUE，2025年5月28日；为减少波动影响，本次测评将相差1分内的模型视为并列。海外产品仅作对比参考，不参与排名。标红为国内前三名。

测评分析

1. o4-mini(high)总分稳居第一。

SuperCLUE评测基准2025年5月总体榜单显示，o4-mini(high)在SuperCLUE-5月测评中表现优异，总分稳居第一，达到70.51分。该模型在推理、代码生成、智能体、指令遵循等多个方面表现出卓越的综合能力，特别是在代码生成）、指令遵循方面得分较高，体现了其强大的逻辑推理和问题解决能力。

2. 国内推理模型崭露头角，部分领域优势突出。

Doubao-1.5-thinking-pro-205415、SenseNova V6 Reasoner等国内模型表现亮眼。其中，Doubao-1.5-thinking-pro-205415在文本创作与理解任务上以81.04的高分领先其他模型。NebulaCoder-V6在数学推理得分上比o4-mini(high)分别高出4.67。

3. 国内大模型在指令遵循方面普遍低于海外模型。

Hunyuan-T1-20250403在国内模型中指令遵循得分第一，为36.97分，但是与海外模型指令遵循得分第一的o4-mini(high)相比，差距达到了31.1分，国内模型在指令遵循方面表现较弱，还有较大的提升空间。

SuperCLUE-基础模型总榜

SuperCLUE测评基准2025年5月基础模型总体表现

排名	模型名称	机构	总分	数学推理	科学推理	代码生成	智能体Agent	精确指令遵循	文本理解与创作	使用方式
🥇	DeepSeek-V3-0324	深度求索	***60.10***	***55.65***	***41.58***	***84.81***	***68.24***	***30.25***	***80.04***	API
-	ChatGPT-4o-latest	OpenAI	***59.58***	49.19	42.08	90.34	67.57	29.13	79.16	API
-	grok-3-latest	X.AI	***57.85***	49.19	34.65	83.04	65.88	34.45	79.90	API
-	Llama-4-Maverick-17B-128E-Instruct	Meta	***52.89***	44.35	38.61	75.15	54.05	30.81	74.39	API
🥈	GLM-4-Air-250414	智谱AI	***51.45***	***53.23***	***35.64***	***77.91***	43.58	***21.85***	76.49	API
🥉	Step-2-16k	阶跃星辰	***44.72***	30.65	18.81	70.81	***49.66***	19.61	***78.76***	API
🥉	Spark4.0 Ultra	科大讯飞	***44.63***	36.29	***22.77***	67.46	44.01	19.89	***77.38***	API
4	MiniMax-Text-01	MiniMax	***43.35***	29.03	20.79	71.20	43.92	19.33	75.80	API
-	Llama-3.3-70B-Instruct	Meta	***42.20***	30.65	12.87	70.22	41.22	27.73	70.51	API
5	Sky-Chat-3.0	昆仑万维	***41.96***	***37.10***	19.80	60.16	35.81	***22.97***	75.91	API
5	kimi-latest	月之暗面	***41.91***	24.19	13.86	***71.99***	***51.69***	15.69	74.04	API
-	Gemma-3-27b-it	Google	***41.74***	32.26	17.82	68.05	34.12	20.45	77.72	API
-	Gemma-3-12b-it	Google	***36.57***	27.42	4.95	68.44	26.69	15.97	75.96	API
-	Gemma-3-4b-it	Google	***27.77***	16.94	2.97	48.92	11.82	11.52	74.47	API
6	GLM-4-9B-Chat	智谱AI	***27.73***	16.94	6.93	43.98	18.24	7.84	72.47	模型
-	Llama-3.1-8B-Instruct	Meta	***24.12***	9.68	4.95	39.25	16.22	8.96	65.64	API
7	MiniCPM3-4B	面壁智能	***21.06***	8.87	4.95	34.91	1.35	7.02	69.28	模型
8	Yi-1.5-9B-Chat-16K	零一万物	***19.81***	6.45	2.97	20.71	12.16	7.56	69.01	模型
-	Mistral-7B-Instruct-v0.3	Mistral AI	***12.47***	5.65	0.99	9.07	1.01	3.36	54.76	模型

注：数据来源SuperCLUE，2025年5月28日；为减少波动影响，本次测评将相差1分内的模型视为并列。海外产品仅作对比参考，不参与排名。标红为国内前三名。

测评分析

1.模型竞争激烈，头部模型优势凸显。

根据基础模型榜单可以发现，DeepSeek-V3-0324以60.10的总分位居榜首，在多个维度展现出强大实力，在科学推理、智能体Agent、文本理解与创作这三个类别都取得了基础模型中的最高分。排名靠前的模型在总分上与靠后模型拉开明显差距，显示出头部模型在综合能力上的优势地位，也反映出大模型领域强者恒强的态势。

2.国内模型在部分领域表现亮眼。

在国内模型表现方面，GLM-4-Air-250414在数学推理上取得53.23的高分，仅次于DeepSeek-V3-0324。这些国内模型在细分领域的出色成绩，表明国内在大模型技术研发上不断取得突破，在特定能力维度上已具备与国际模型竞争的实力。

3.各维度能力发展不均衡。

从各模型在不同维度的得分来看，能力发展不均衡现象明显。如代码生成维度，DeepSeek-V3-0324得分84.81，而部分模型得分较低，差距巨大。在精确指令遵循维度，模型间分数差异也较为显著。这种不均衡体现了不同模型在能力侧重上的差异，也反映出大模型在追求综合能力提升时，仍面临各维度能力协调发展的挑战。

开源模型榜单

SuperCLUE开源榜单

排名	模型名称	机构	参数量	总分
1	DeepSeek-R1	深度求索	671B	*61.94*
2	DeepSeek-V3-0324	深度求索	671B	*60.10*
2	Qwen3-32B(Thinking)	阿里巴巴	32B	*59.32*
3	Qwen3-235B-A22B(Thinking)	阿里巴巴	235B	*59.00*
3	Pangu Pro MoE-72B-A16B	华为	72B	*58.75*
4	Qwen3-14B(Thinking)	阿里巴巴	14B	*57.12*
5	Qwen3-8B(Thinking)	阿里巴巴	8B	*53.78*
5	Qwen3-30B-A3B(Thinking)	阿里巴巴	30B	*53.27*
-	Llama-4-Maverick-17B-128E-Instruct	Meta	400B	*52.89*
6	GLM-Z1-AirX	智谱AI	32B	*51.55*
6	GLM-4-Air-250414	智谱AI	32B	*51.45*
6	DeepSeek-R1-Distill-Qwen-32B	深度求索	32B	*50.81*
7	Qwen3-4B(Thinking)	阿里巴巴	4B	*46.04*
8	MiniMax-Text-01	MiniMax	456B	*43.35*
-	Llama-3.3-70B-Instruct	Meta	70B	*42.20*
-	Gemma-3-27b-it	Google	27B	*41.74*
-	Gemma-3-12b-it	Google	12B	*36.57*
-	Gemma-3-4b-it	Google	4B	*27.77*
9	GLM-4-9B-Chat	智谱AI	9B	*27.73*
-	Llama-3.1-8B-Instruct	Meta	8B	*24.12*
10	MiniCPM3-4B	面壁智能	4B	*21.06*
11	Yi-1.5-9B-Chat-16K	零一万物	9B	*19.81*
-	Mistral-7B-Instruct-v0.3	Mistral AI	7B	*12.47*

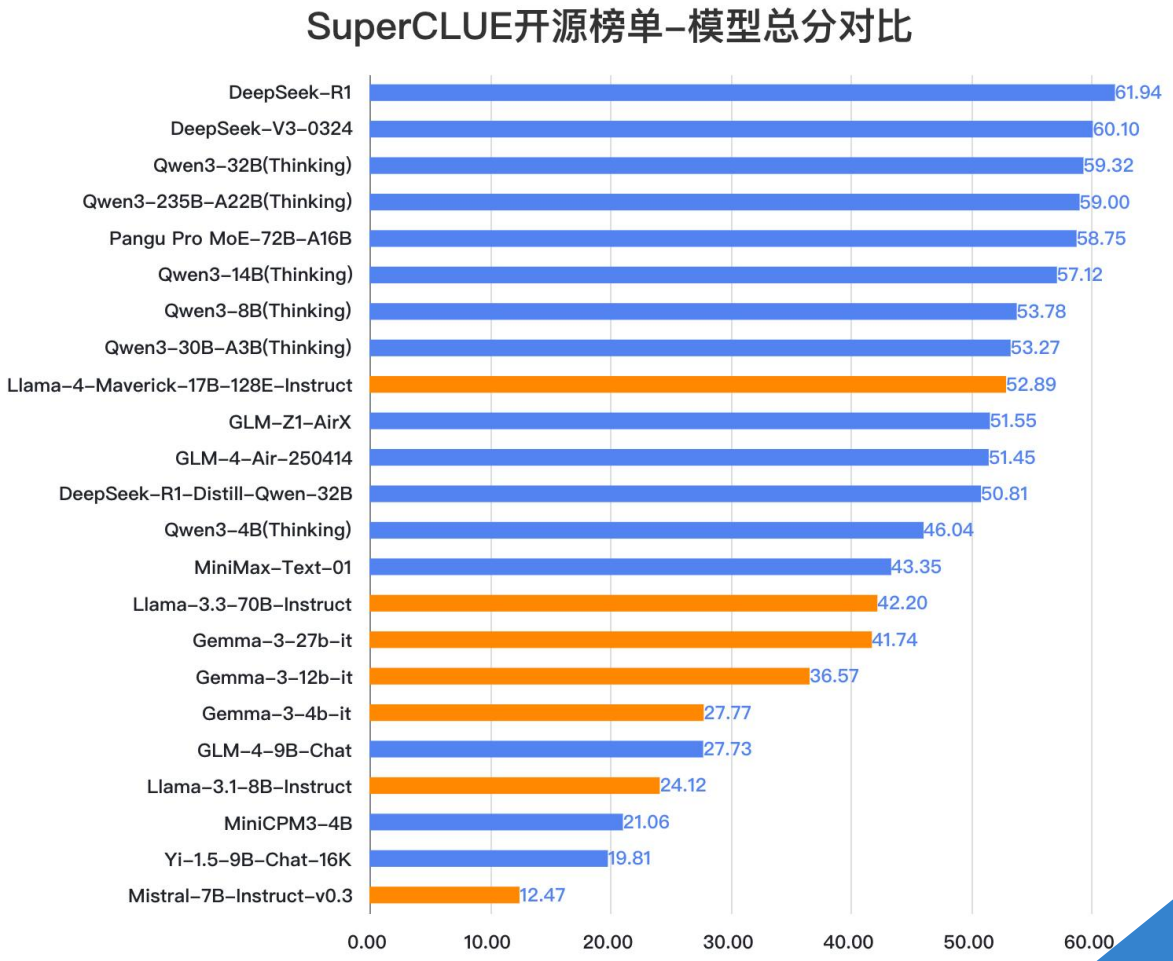
数据来源：SuperCLUE，2025年5月28日；

注：由于部分模型分数较为接近，为了减少问题波动对排名的影响，本次测评将相距1分区间的模型定义为并列。其中模型参数量数据来源于官方披露，若模型为 MoE 架构，以总参数量为准。

开源模型分析

✓ 中文场景下，国内开源模型已具备较大优势

DeepSeek系列开源模型、Qwen系列开源模型，在5月SuperCLUE测评中表现优异，均有超过Llama-4-Maverick-17B-128E-Instruct的表现，引领全球开源生态。



SuperCLUE-推理模型总榜

SuperCLUE测评基准2025年5月推理模型表现

排名	模型名称	机构	推理榜单总分	数学推理	科学推理	代码生成	使用方式
1	Doubao-1.5-thinking-pro-250415	字节跳动	***67.4***	62.10	***52.08***	***87.97***	API
1	NebulaCoder-V6	中兴通讯	***67.4***	***62.39***	***53.19***	86.59	API
-	o4-mini(high)	OpenAI	***67.1***	57.72	52.00	91.52	API
1	360zhinao2-o1.5	360	***66.8***	***62.86***	50.52	87.14	API
2	DeepSeek-R1	深度求索	***66.1***	62.10	49.50	86.59	API
2	Qwen3-235B-A22B(Thinking)	阿里巴巴	***66.0***	58.87	48.51	***90.53***	API
2	SenseNova V6 Reasoner	商汤	***65.3***	62.10	47.52	86.19	API
3	Qwen3-30B-A3B(Thinking)	阿里巴巴	***64.9***	61.29	50.50	83.04	API
-	Gemini 2.5 Flash Preview 04-17	Google	***63.8***	57.26	47.52	86.59	API
4	Qwen3-32B(Thinking)	阿里巴巴	***63.7***	56.45	47.52	***87.18***	API
-	Gemini 2.5 Pro Preview 05-06	Google	***63.5***	55.65	43.56	91.32	API
-	Gemini 2.5 Flash Preview 05-20	Google	***63.3***	56.45	46.53	86.79	API
4	ERNIE-X1-Turbo-32K	百度	***63.0***	57.63	48.98	82.25	API
4	Hunyuan-T1-20250403	腾讯	***62.9***	57.50	47.52	83.63	API
5	Qwen3-14B(Thinking)	阿里巴巴	***62.1***	57.26	***51.49***	77.51	API
-	Claude-Opus-4-Reasoning	Anthropic	***61.4***	51.61	45.54	86.98	POE
6	GLM-Z1-AirX	智谱AI	***60.8***	***64.41***	46.00	71.99	API
6	Pangu Pro MoE-72B-A16B	华为	***60.3***	59.68	39.60	81.66	API
7	Qwen3-8B(Thinking)	阿里巴巴	***58.5***	60.48	42.57	72.39	API
7	DeepSeek-R1-Distill-Qwen-32B	深度求索	***58.3***	60.33	38.89	75.54	API
8	kimi-K1.5	月之暗面	***57.2***	59.68	38.61	73.37	网页
-	o3	OpenAI	***57.0***	56.45	25.74	88.76	POE
9	Spark X1	科大讯飞	***53.0***	53.23	31.68	73.96	API
9	Qwen3-4B(Thinking)	阿里巴巴	***52.8***	50.81	33.66	73.77	API

注：数据来源SuperCLUE，2025年5月28日；为减少波动影响，本次测评将相差1分内的模型视为并列。海外产品仅作对比参考，不参与排名。

测评分析

1.国内模型表现亮眼，竞争激烈。

根据推理模型榜单可以发现，国内模型成绩突出。NebulaCoder-V6、Doubao-1.5-thinking-pro-250415和360zhinao2-o1.5并列第一，DeepSeek-R1、Qwen3-235B-A22B(Thinking)和SenseNova V6 Reasoner并列国内推理模型第二。国内模型在数学推理、科学推理、代码生成三大任务上相互竞争，彰显出国内模型发展的蓬勃态势。

2.数学推理与科学推理能力分化明显，头部模型优势突出。

在数学推理维度，GLM-Z1-AirX以64.41分领跑；而科学推理方面，NebulaCoder-V6、Doubao-1.5-thinking-pro-250415领先于其他模型。反观排名较后的模型，如Spark X1在科学推理上差距显著，反映出模型在跨学科推理能力上的不均衡发展。

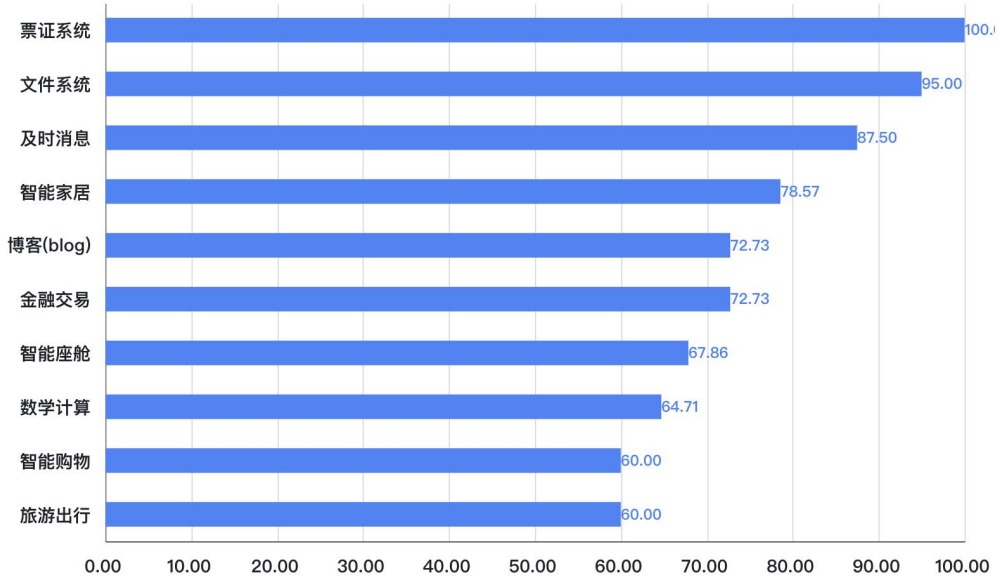
3.代码生成能力成为竞争新焦点，国内模型实现突破。

在代码生成任务中，Qwen3-235B-A22B(Thinking)以90.53分位居国内榜首，不仅在国内领先，更与OpenAI的o4-mini(high)差距微小。

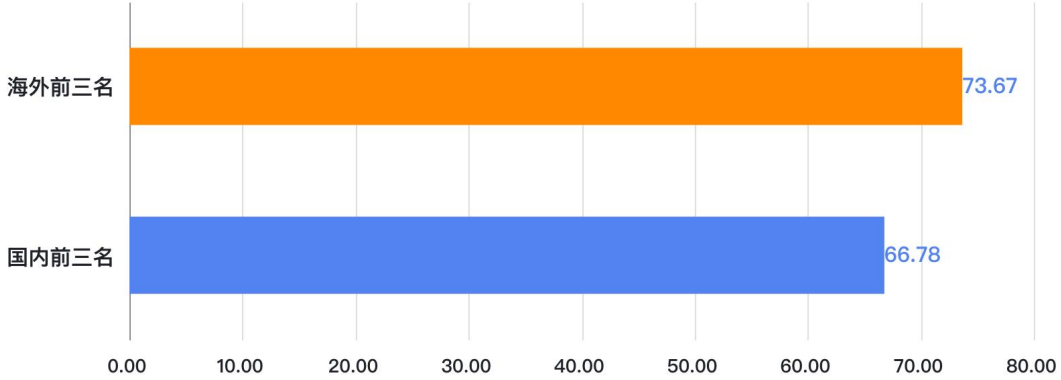
SuperCLUE-Agent：智能体测评分析

SuperCLUE-Agent

各应用场景下模型最高得分



国内外前三名平均得分



数据来源：SuperCLUE，2025年5月28日。

测评分析

1.各应用场景上成熟度差异显著

统计每个场景下的最高得分，发现在“票证系统”和“文件系统”场景下模型最高得分分别为：100分和95分，成熟度较高；而智能购物和旅游出行两个场景的模型最高得分只有60分，成熟度较低。

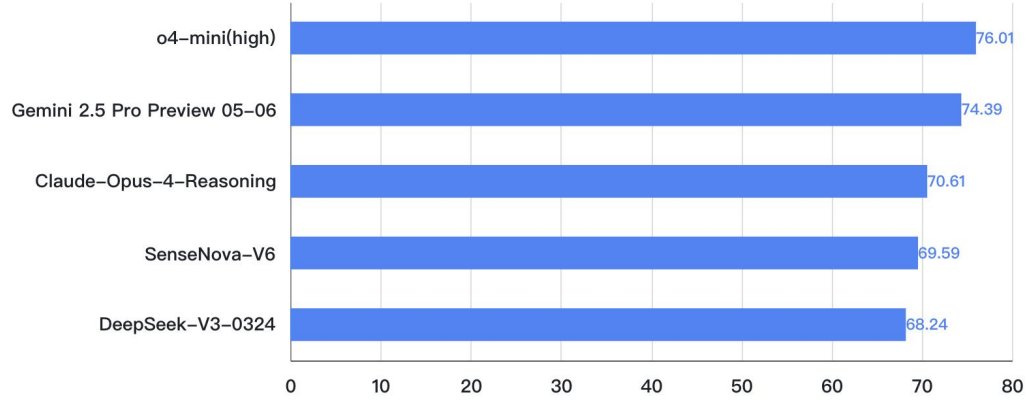
2.海外模型领先，国内最好模型与海外顶尖水平仍有一定差距。

海外最高为o4-mini(high)，得分为76.01，比国内最高SenseNova V6 Reasoner(69.59)高出6.42分。国内前三名的平均成绩（66.78）相比于海外前三名的平均分（73.67）低6.89分。

3.推理模型领先于基础模型。

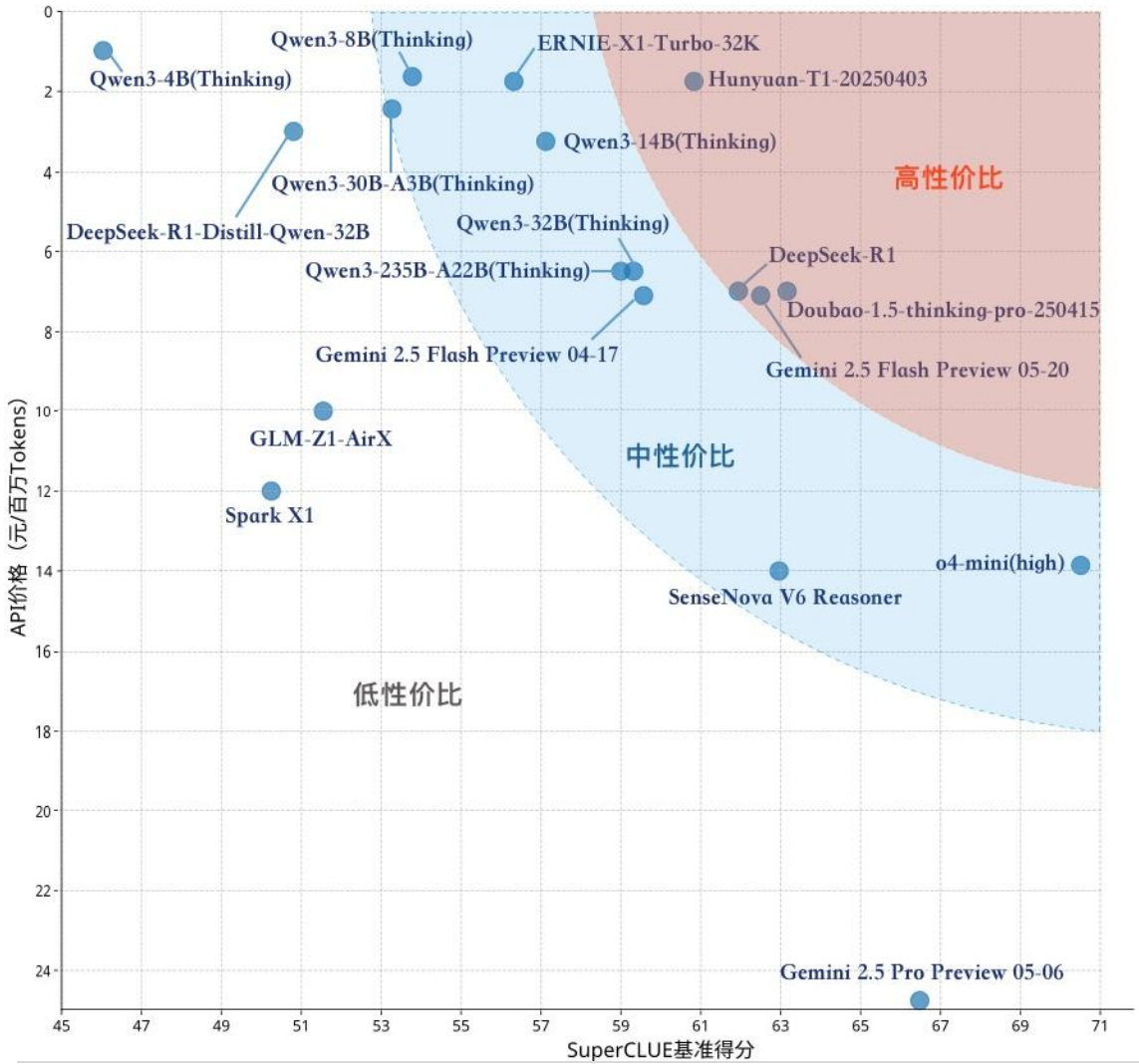
表现最好的推理模型o4-mini(high)得分比表现最好的基础模型DeepSeek-V3-0324高7.77分。排名前五的模型中有4个为推理模型。

智能体Agent任务前5名模型得分



SuperCLUE性价比区间分布-推理模型

推理模型性价比分布



趋势分析

1. 国产推理模型凭借较低的价格实现高质量输出，展现出显著的性价比优势。

国产推理模型Doubao-1.5-thinking-pro-250415和DeepSeek-R1在性价比方面展现出强大竞争力。海外模型Gemini 2.5 Flash Preview 05-20也具备高性价比，国产模型Hunyuan-T1-20250403性价比也很高，虽然得分相比于之前三款模型较低，但是价格便宜很多。

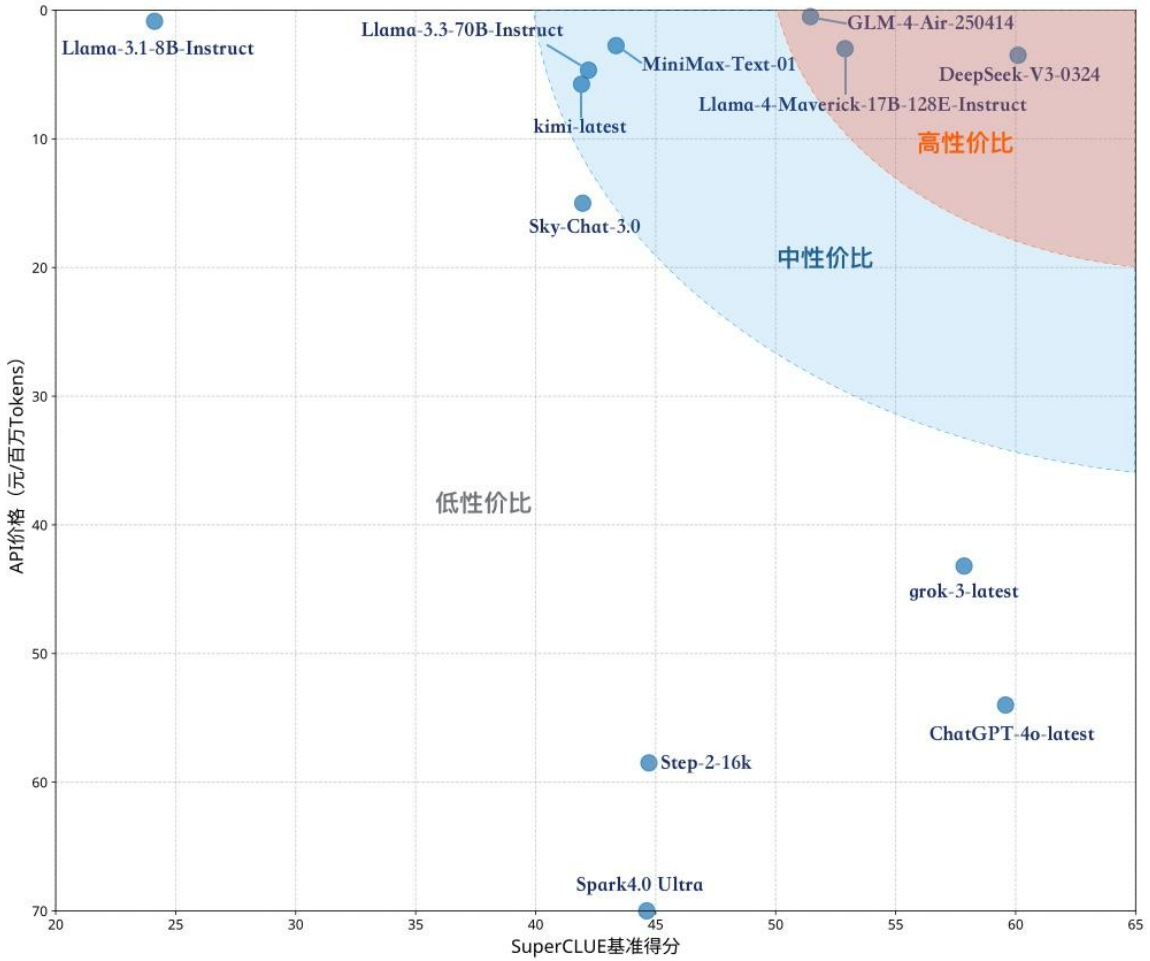
2.推理模型的能力与其API价格无关，o4-mini(high)以中等的价位领先其他模型。

Gemini 2.5 Pro Preview 05-06得分排在第二，但是价格远远高于其他模型；得分较高的推理模型大多集中在中高性价比区域。o4-mini(high)以中等的价位领先其他模型。

数据来源：SuperCLUE，2025年5月28日；开源模型如Qwen3-32B(Thinking)使用方式为API，价格信息均来自官方信息。
注：部分模型API的价格是分别基于输入和输出的 tokens 数量确定的。这里我们依照输入 tokens 与输出 tokens 3:1 的比例来估算其整体价格。价格信息取自官方在5月的标准价格（非优惠价格）。

SuperCLUE性价比区间分布-基础模型

基础模型性价比分布



趋势分析

1. 国产基础模型凭借较低的价格实现高质量输出，展现出显著的性价比优势。

国产基础模型DeepSeek-V3-0324在性价比方面展现出强大竞争力，以最高得分领先其他模型，同时又保持了极低的应用成本。国产模型GLM-4-Air-250414和海外模型Llama-4-Maverick-17B-128E-Instruct也具有高性价比。

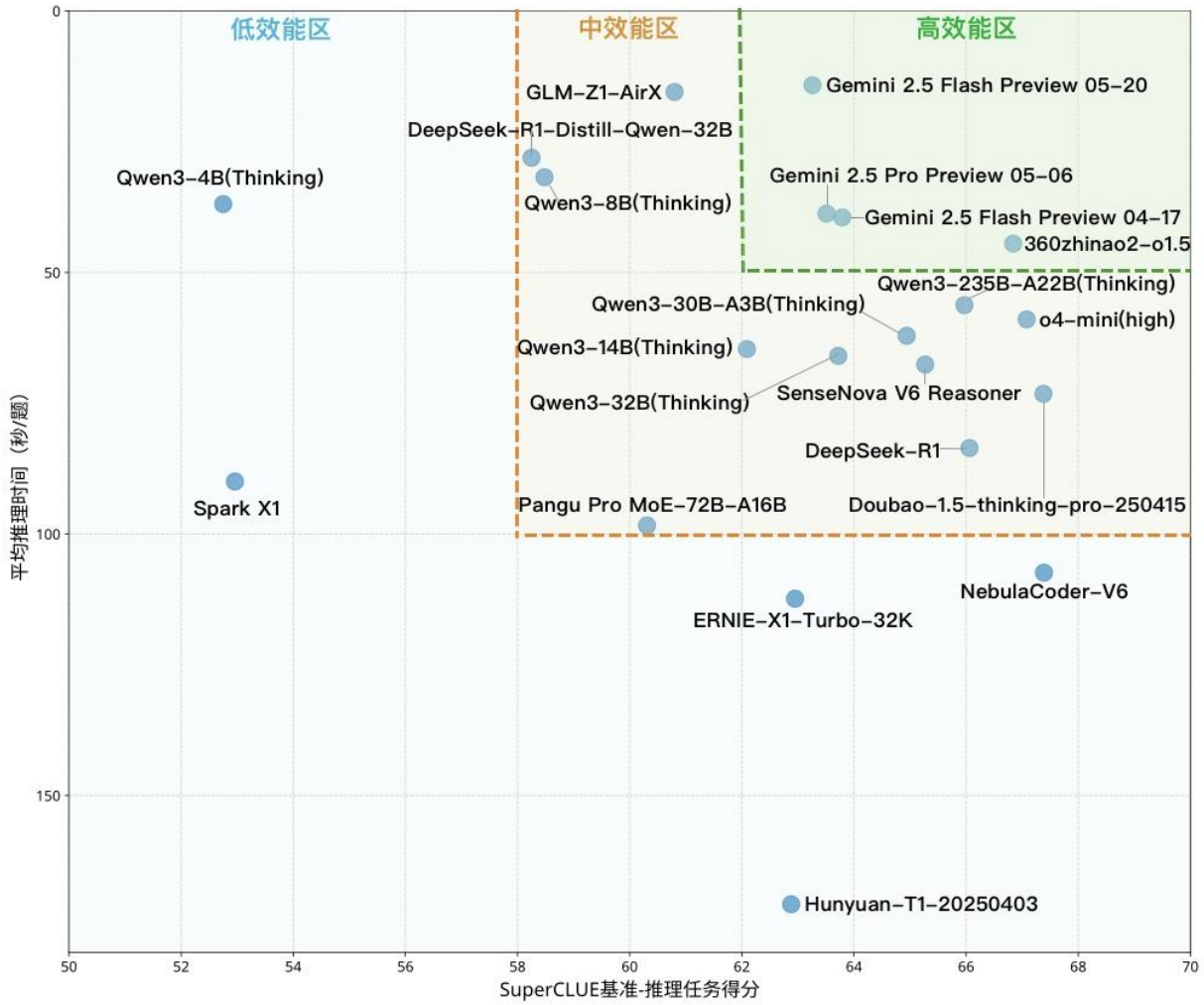
2. 基础模型的综合能力与其API价格无关。

DeepSeek-V3-0324以低价格最高得分领先其他模型。ChatGPT-4o-latest得分排在第二，但是价格远远高于其他模型。高性价比和中性价比的模型价格都在10元/百万Tokens之内。

数据来源：SuperCLUE，2025年5月28日；开源模型如Qwen3-32B(Thinking)使用方式为API，价格信息均来自官方信息。
注：部分模型API的价格是分别基于输入和输出的 tokens 数量确定的。这里我们依照输入 tokens 与输出 tokens 3:1 的比例来估算其整体价格。价格信息取自官方在5月的标准价格（非优惠价格）。

SuperCLUE大模型综合效能区间分布-推理模型

推理模型推理效能区间



数据来源：SuperCLUE，2025年5月28日；
模型推理速度选取5月测评中具有公开API的模型。平均推理时间为所有测评数据推理时间的平均值（秒）。
推理任务得分为推理任务总分：数学推理、科学推理和代码生成的平均分。

趋势分析

1.360zhinao2-o1.5和Gemini系列模型综合效能表现领先。

本次测评有4个模型处于高效能区间，Gemini系列模型就占了3个，360zhinao2-o1.5在准确度方面高于Gemini，但在推理速度方面不如Gemini。

2.国内推理模型大部分处于中低效能区间。

除360zhinao2-o1.5外，国内推理模型皆处于中低效能区间，如Qwen3系列推理模型绝大多数都处于中效能区，在推理任务上得分可以比肩Gemini系列模型，但在推理速度方面仍有较大提升空间。

3.兼顾推理速度和准确度是重难点。

对于在推理任务上得分较高的一系列模型，如NebulaCoder-V6、Doubao-1.5-thinking-pro-250415等，其平均推理时间也达到了60秒以上。为了更好地将推理模型应用于实际场景，需要考虑如何在确保较高推理准确度的同时提升推理速度。

SuperCLUE大模型综合效能区间分布-基础模型

基础模型推理效能区间

数据来源：SuperCLUE，2025年5月28日；
模型推理速度选取5月测评中具有公开API的模型。平均推理时间为所有测评数据推理时间的平均值（秒）。
总分为六大任务的平均分。

趋势分析

1.海外基础模型综合效能领先。

ChatGPT-4o-latest和grok-3-latest处于高效能区间，推理速度和准确度都表现优异。在推理速度上，Llama-4-Maverick-17B-128E-Instruct以唯一一个推理耗时在10秒以内领先其他模型。

2.国内基础模型大部分处于中低效能区间。

国内基础模型中，GLM-4-Air-250414是唯一一个处于高效能区间的模型，推理速度和准确度相较于高区间的其他模型都较差。

3.基础模型在提升准确度和减少推理时间方面有提升空间。

被测的基础模型推理耗时均在10秒到20秒之间，但得分都在60分以下，仍有一定的提升空间。

国内大模型成熟度-SC成熟度指数

SuperCLUE大模型能力成熟度指数-SC指数

指数序号	能力	最高分	最低分	成熟度SC指数	成熟度区间
1	文本理解与创作	81.04	74.04	0.91	高成熟度 (>0.8)
2	代码生成	87.97	60.16	0.68	中成熟度 (0.5-0.8)
3	智能体Agent	69.59	35.81	0.51	
4	精确指令遵循	36.97	15.69	0.42	低成熟度 (0.2-0.5)
5	数学推理	62.86	24.19	0.38	
6	科学推理	53.19	13.86	0.26	

数据来源：SuperCLUE，2025年5月28日；SC成熟度指数=国内闭源模型最差成绩/国内闭源模型最好成绩。

国内大模型成熟度分析

1.高成熟度能力

- ✓ 高成熟度指大部分闭源大模型普遍擅长的能力，SC成熟度指数在0.8至1.0之间。
- ✓ 当前国内大模型成熟度较高的能力是【文本理解与创作】，也是目前产业和用户侧大模型的重点应用场景。

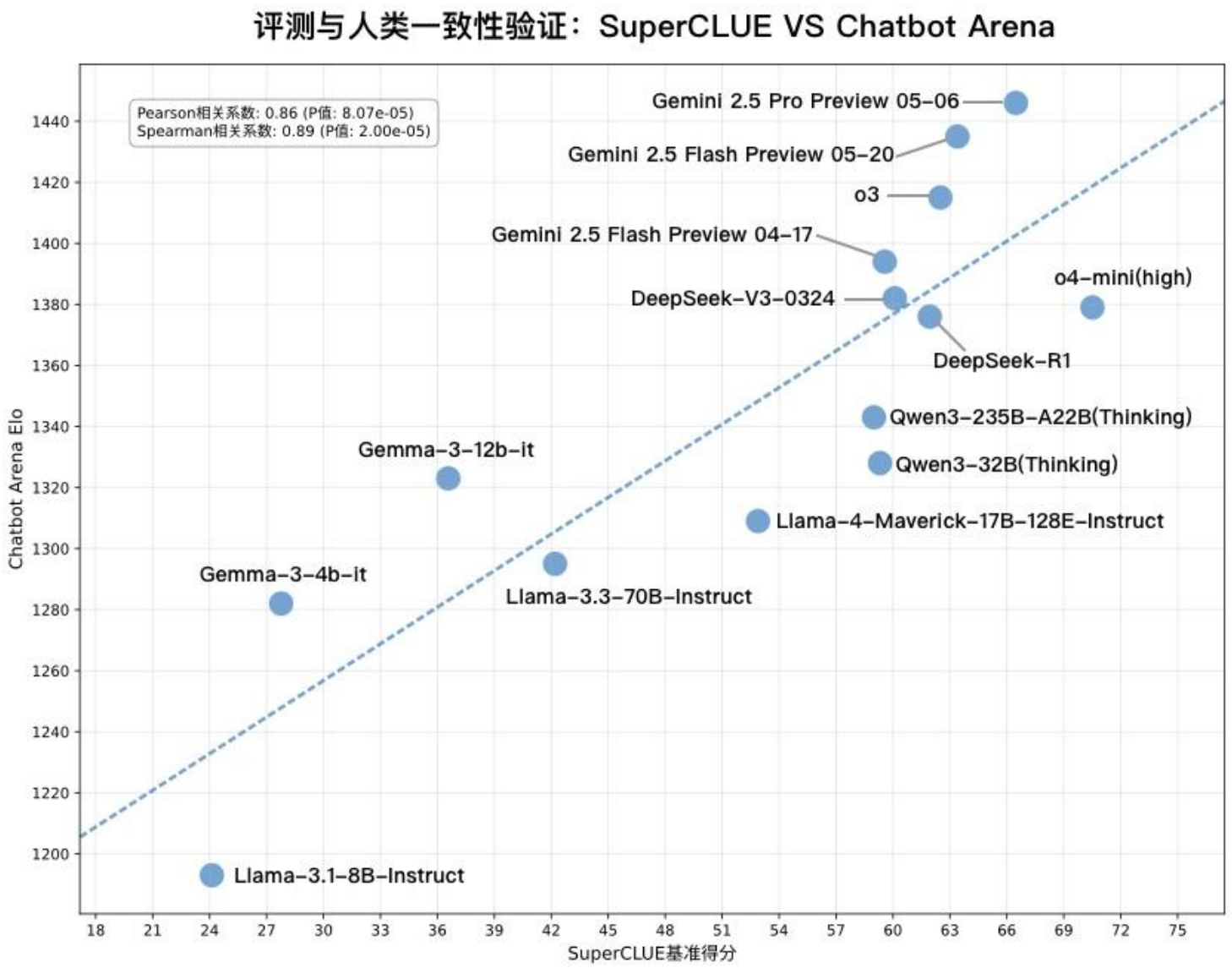
2.中成熟度能力

- ✓ 中成熟度指的是不同大模型能力上有一定区分度，但不会特别大。SC成熟度指数在0.5至0.8之间。
- ✓ 当前国内大模型表现出中成熟度的能力是【代码生成】、【智能体Agent】，还有一定优化空间。

3.低成熟度能力

- ✓ 低成熟度指的是少量大模型较为擅长，很多模型无法胜任。SC成熟度指数在0.2至0.5之间。
- ✓ 当前国内大模型低成熟度的能力是【精确指令遵循】、【数学推理】和【科学推理】。

评测与人类一致性验证：对比Chatbot Arena



数据来源：SuperCLUE，2025年5月28日。

Chatbot Arena是当前英文领域较为权威的大模型排行榜，由LMSYS Org开放组织构建，它以公众匿名投票的方式，对各种大型语言模型进行对抗评测。

将SuperCLUE得分与ChatBot Arena得分进行相关性计算，得到：

皮尔逊相关系数：0.86，P值：8.07e-05；

斯皮尔曼相关系数：0.89，P值：2.00e-05。

说明SuperCLUE基准测评的成绩，与人类对模型的评估（以大众匿名投票的Chatbot Arena为典型代表），具有高度一致性。

SuperCLUE-推理任务总榜

SuperCLUE测评基准2025年5月推理任务表现

排名	模型名称	机构	推理榜单总分	数学推理	科学推理	代码	使用方式
🥇	Doubao-1.5-thinking-pro-250415	字节跳动	***67.4***	62.10	***52.08***	***87.97***	API
🥇	NebulaCoder-V6	中兴	***67.4***	***62.39***	***53.19***	86.59	API
-	o4-mini(high)	OpenAI	***67.1***	57.72	52.00	91.52	API
🥇	360zhinao2-o1.5	360	***66.8***	***62.86***	50.52	87.14	API
🥈	DeepSeek-R1	深度求索	***66.1***	62.10	49.50	86.59	API
🥈	Qwen3-235B-A22B(Thinking)	阿里巴巴	***66.0***	58.87	48.51	***90.53***	API
🥈	SenseNova V6 Reasoner	商汤	***65.3***	62.10	47.52	86.19	API
🥉	Qwen3-30B-A3B(Thinking)	阿里巴巴	***64.9***	61.29	50.50	83.04	API
-	Gemini 2.5 Flash Preview 04-17	Google	***63.8***	57.26	47.52	86.59	API
4	Qwen3-32B(Thinking)	阿里巴巴	***63.7***	56.45	47.52	***87.18***	API
-	Gemini 2.5 Pro Preview 05-06	Google	***63.5***	55.65	43.56	91.32	API
-	Gemini 2.5 Flash Preview 05-20	Google	***63.3***	56.45	46.53	86.79	API
4	ERNIE-X1-Turbo-32K	百度	***63.0***	57.63	48.98	82.25	API
4	Hunyuan-T1-20250403	腾讯	***62.9***	57.50	47.52	83.63	API
5	Qwen3-14B(Thinking)	阿里巴巴	***62.1***	57.26	***51.49***	77.51	API
-	Claude-Opus-4-Reasoning	Anthropic	***61.4***	51.61	45.54	86.98	POE
6	GLM-Z1-AirX	智谱AI	***60.8***	***64.41***	46.00	71.99	API
6	DeepSeek-V3-0324	深度求索	***60.7***	55.65	41.58	84.81	API
-	ChatGPT-4o-latest	OpenAI	***60.5***	49.19	42.08	90.34	API
6	Pangu Pro MoE-72B-A16B	华为	***60.3***	59.68	39.60	81.66	API
7	Qwen3-8B(Thinking)	阿里巴巴	***58.5***	60.48	42.57	72.39	API
7	DeepSeek-R1-Distill-Qwen-32B	深度求索	***58.3***	60.33	38.89	75.54	API
8	kimi-K1.5	月之暗面	***57.2***	59.68	38.61	73.37	网页
-	o3	OpenAI	***57.0***	56.45	25.74	88.76	POE
-	grok-3-latest	X.AI	***55.6***	49.19	34.65	83.04	API
9	GLM-4-Air-250414	智谱AI	***55.6***	53.23	35.64	77.91	API
10	Spark X1	科大讯飞	***53.0***	53.23	31.68	73.96	API
10	Qwen3-4B(Thinking)	阿里巴巴	***52.7***	50.81	33.66	73.77	API
-	Llama-4-Maverick-17B-128E-Instruct	Meta	***52.7***	44.35	38.61	75.15	API
11	Spark4.0 Ultra	科大讯飞	***42.2***	36.29	22.77	67.46	API
12	MiniMax-Text-01	MiniMax	***40.3***	29.03	20.79	71.20	API
12	Step-2-16k	阶跃星辰	***40.1***	30.65	18.81	70.81	API
-	Gemma-3-27b-it	Google	***39.4***	32.26	17.82	68.05	API
13	Sky-Chat-3.0	昆仑万维	***39.0***	37.10	19.80	60.16	API
-	Llama-3.3-70B-Instruct	Meta	***37.9***	30.65	12.87	70.22	API
14	kimi-latest	月之暗面	***36.7***	24.19	13.86	71.99	API
-	Gemma-3-12b-it	Google	***33.6***	27.42	4.95	68.44	API
-	Gemma-3-4b-it	Google	***22.9***	16.94	2.97	48.92	API
15	GLM-4-9B-Chat	智谱AI	***22.6***	16.94	6.93	43.98	模型
-	Llama-3.1-8B-Instruct	Meta	***18.0***	9.68	4.95	39.25	API
16	MiniCPM3-4B	面壁智能	***16.2***	8.87	4.95	34.91	模型
17	Yi-1.5-9B-Chat-16K	零一万物	***10.0***	6.45	2.97	20.71	模型
-	Mistral-7B-Instruct-v0.3	Mistral AI	***5.2***	5.65	0.99	9.07	模型

注：数据来源SuperCLUE，2025年5月28日；为减少波动影响，本次测评将相差1分内的模型视为并列。海外产品仅作对比参考，不参与排名。标红为国内前三名。

测评分析

1. 国内模型在推理任务中竞争力强劲，头部模型得分紧密。

从榜单来看，前几名模型总分差距微小，Doubao-1.5-thinking-pro-250415等并列排名第一的模型与并列排名第二的SenseNova V6 Reasoner等模型总分极为接近，显示推理能力的高水准竞技。国内模型表现突出，标红的国内模型前三占据前列，与国外OpenAI、Google等的部分模型形成有力竞争。

2. 推理子任务分化明显，模型优势场景各异。

各推理子任务中，模型表现呈现差异化：数学推理以GLM-Z1-AirX为代表，展现强大的数学逻辑推导能力；科学推理中，NebulaCoder-V6和Doubao-1.5-thinking-pro-250415得分领先，擅长科学问题分析；代码生成任务里，Qwen3-235B-A22B（90.53）和Doubao-1.5-thinking-pro-250415（87.97）表现优异，凸显编程推理优势。这种分化表明模型针对不同场景（如数学计算、科学研究、代码开发）进行了专项优化，用户可根据具体需求选择适配模型，实现推理能力的精准应用，也为模型在垂直领域的深度拓展提供了方向。

10B级别小模型榜单

SuperCLUE-10B级别小模型榜单

排名	模型名称	机构	参数量	总分
1	Qwen3-8B(Thinking)	阿里巴巴	8B	*53.78*
2	Qwen3-4B(Thinking)	阿里巴巴	4B	*46.04*
-	Gemma-3-4b-it	Google	4B	*27.77*
3	GLM-4-9B-Chat	智谱AI	9B	*27.73*
-	Llama-3.1-8B-Instruct	Meta	8B	*24.12*
4	MiniCPM3-4B	面壁智能	4B	*21.06*
5	Yi-1.5-9B-Chat-16K	零一万物	9B	*19.81*
-	Mistral-7B-Instruct-v0.3	Mistral AI	7B	*12.47*

10B级别小模型分析

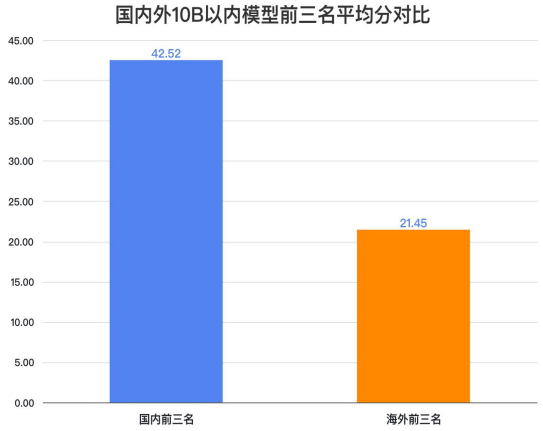
✓ **10B级别模型中，DeepSeek-R1-Distill-Qwen-7B和Gemma-2-9b-it分列国内外榜首**

在本次SuperCLUE测评中，Qwen3-8B(Thinking)取得53.78分，取得10B以内模型的最高分。Qwen3-4B(Thinking)取得46.04分排名国内第2，GLM-4-9B-Chat以27.73分位于国内第三。Gemma-3-4b-it取得27.77分，领跑海外10B以内模型。

✓ **国内10B以内模型进展迅速，展现出极致的性价比**

在10B以内模型中，超过40分的模型有2个，分别为Qwen3系列的Qwen3-8B(Thinking)和Qwen3-4B(Thinking)，均是国内大模型。展现出10B以内小参数量级模型的极致的性价比。

国内外对比



➢ **国内头部10B以内模型平均水平领先于海外模型**

✓ 从国内外10B小模型能力的对比数据看，国内10B小模型优势较大。**国内10B小模型前三名模型的得分相较于国外前三名平均高出21.07分。**

数据来源：SuperCLUE，2025年5月28日；
注：由于部分模型分数较为接近，为了减少问题波动对排名的影响，本次测评将相距1分区间的模型定义为并列。其中模型参数量数据来源于官方披露，若模型为 MoE 架构，以总参数量为准。

端侧5B级别小模型榜单

2025年端侧小模型快速发展，已在设备端侧（非云）上实现本地运行，其中PC、手机、智能眼镜、机器人等大量场景已展现出极高的落地可行性。

➢ **国内端侧小模型进展迅速，相比国外小模型，国内小模型在中文场景下展现出更好的性能表现**

✓ Qwen3-4B(Thinking)表现惊艳，取得总分46.04分的优异成绩，在SuperCLUE端侧5B小模型榜单中排名榜首。其中文本理解与创作78.50分，与同等参数量级模型Gemma-3-4b-it相比多个维度均有不同幅度的领先，展示出小参数量级模型极高的性价比。

✓ MiniCPM3-4B小模型同样表现不俗，取得总分21.06分，文本理解与创作的成绩已接近70分。

SuperCLUE端侧5B级别小模型榜单

排名	模型名称	机构	参数量	总分	数学推理	科学推理	代码生成	智能体Agent	精确指令遵循	文本理解与创作	使用方式
1	Qwen3-4B(Thinking)	阿里巴巴	4B	46.04	50.81	33.66	73.77	17.91	21.57	78.50	API
-	Gemma-3-4b-it	Google	4B	27.77	16.94	2.97	48.92	11.82	11.52	74.47	API
2	MiniCPM3-4B	面壁智能	4B	21.06	8.87	4.95	34.91	1.35	7.02	69.28	模型

数据来源：SuperCLUE，2025年5月28日。

联系我们

——立足业内领先的第三方大模型测评机构，致力于为业界提供专业测评服务：

通用大模型测评

提供大模型综合性评测服务 ，输出全方位的评测报告，包括但不限于多维度测评结果、横向对比、典型示例、模型优化建议。

行业与专项大模型测评

聚焦测评大模型在行业落地应用效果，包括但不限于汽车、手机、金融、工业、教育、医疗等行业大模型应用能力，中文Agent能力测评、大模型安全评估、多模态能力测评、个性化角色扮演能力测评。

多模态大模型测评

多维度全方位测评多模态大模型的基础能力与应用能力，包括但不限于实时多模态交互、视频生成基准测评、文生图测评、多模态理解测评等。

AI应用测评

提供AI大模型落地应用及工具测评，包括但不限于生产力工具、代码助手、AI搜索等应用；AI PC、AI手机、XR设备及具身智能等设备端应用。

大模型深度研究报告

提供国内外大模型深度研究报告，全面调研与分析国内外大模型技术进展及应用落地情况 ，为企事业单位提供及时、深度的第三方专业报告。

业务合作：请简要描述需求至合作邮箱 contact@superclue.ai

SuperCLUE

交流
合作

扫码
关注

- 排行榜官方地址：https://www.superclueai.com
- 官网：www.CLUEbenchmarks.com
- Github地址：https://github.com/CLUEbenchmark
- 联系人：徐老师 18806712650（微信同号）　朱老师 18621237819（微信同号）

法律声明

- **版权声明**

本报告为SuperCLUE团队制作，其版权归属SuperCLUE，任何机构和个人引用或转载本报告时需注明来源为SuperCLUE，且不得对本报告进行任何有悖原意的引用、删节和修改。任何未注明出处的引用、转载和其他相关商业行为都将违反《中华人民共和国著作权法》和其他法律法规以及有关国际公约的规定。对任何有悖原意的曲解、恶意解读、删节和修改等行为所造成的一切后果，SuperCLUE不承担任何法律责任，并保留追究相关责任的权力。

- **免责条款**

本报告基于中文大模型基准测评（SuperCLUE）5月的自动化测评结果以及已公开的信息编制，力求结果的真实性和客观性。然而，所有数据和分析均基于报告出具当日的情况，对未来信息的持续适用性或变更不承担保证。本报告所载的意见、评估及预测仅为出具日的观点和判断，且在未来无需通知即可随时更改。可能根据不同假设、研究方法、即时动态信息和市场表现，发布与本报告不同的意见、观点及预测，无义务向所有接受者进行更新。

本团队力求报告内容客观、公正，但本报告所载观点、结论和建议仅供参考使用，不作为投资建议。对依据或者使用本报告及本公司其他相关研究报告所造成的一切后果，本公司及作者不承担任何法律责任。